COMPREHENSIVE CHARACTERIZATION OF MAIZE LANDRACES: INTEGRATIVE STRATEGIES TO IDENTIFY AND DEPLOY USEFUL ALLELIC DIVERSITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jorge Alberto Romero Navarro

August 2017

COMPREHENSIVE CHARACTERIZATION OF MAIZE LANDRACES: INTEGRATIVE STRATEGIES TO IDENTIFY AND DEPLOY USEFUL ALLELIC DIVERSITY

Jorge Alberto Romero Navarro, Ph.D.

Cornell University 2017

The characterization of natural genetic diversity and the exploration of its relationship with variation in phenotypic traits is of great interest for evolutionary, conservation, and improvement purposes; in addition, understanding the relationships between genotype, phenotype, and the environment can provide insight on the molecular pathways controlling quantitative traits, as well as their fitness implications. The following work entails the characterization of a comprehensive panel of maize landraces from Latin America and is divided in 5 sections.

The chapter "Identifying the diamond in the rough: a study of allelic diversity underlying flowering time adaptation in maize landraces" describes the general experimental design used to characterize the landrace panel. In this chapter the relationship between large scale altitude and latitude adaptation is also explored, and the trait flowering time is used as a case study to explore the genetic architecture of a complex trait through field experiments. The chapter "Genome-environment association allows identifying useful adaptive alleles from maize landraces" explores further the relationship between the genotypic variation in landraces and their adaptation to local abiotic environmental conditions. Because landraces have evolved for thousands of years in those environments, we observe significant association at candidate genes and observed

that adaptive alleles are common and shared across populations, which has important consequences for future breeding efforts. The chapter ”Genome wide association for plant height variation in maize landraces” represents the analysis of the genetic basis of plant height variation in landraces. Plant height, like yield, is a very complex trait with a significant heritable component. The association at key hormonal regulators, as well as flowering time associated regions, shows the potential to unveil genes underlying this trait, however the results of phenotypic prediction suggest that higher marker density is necessary to study traits on this order of complexity in a panel of very diverse landraces. The chapter ”Exploring the potential for finding sources of resistance to Fusarium ear rot among maize landraces” represents the analyses of phenotypic evaluation of inoculated ear rot trials. Lastly, the final chapter ”Integration of controlled populations and association mapping to score cytological features in the maize genome” describes the joint analyses of a mapping population segregating for the abnormal chromosome 10 and the landraces accessions. By combining the results of both populations, putative calls are made in the landraces regarding their chromosome 10 allele.

BIOGRAPHICAL SKETCH

Alberto was born in Mexico City, Mexico. From a young age, he was interested in science and plants, showing great potential for the first and difficulty growing the second. Starting at age 16 he became involved in scientific research through a program called Youth for Research, directed by the National Autonomous University of Mexico (UNAM) in Mexico City. As part of this program, he was involved in two research projects. The first project, supervised by Dr Victor Chavez, entailed the use of plant tissue culture for the propagation of endemic species, which helped him better understand plant physiology and growth, and the second, supervised by Professor Edda Sciutto, involved research on the properties of a chimeric protein as a delivery mechanism for a vaccine against the parasite worm *Taenia solium*. Together, both projects showed him how science was improving both environment conservation and human health. After his experience in those internships, Alberto decided to pursue a degree with a significant research focus, and completed his Bachelor of Science studying Genomic Sciences at UNAM's Center for Genomic Sciences in Morelos, Mexico. While a student, he returned to his roots, doing a research project in plants, in this case using as model organism the common bean *Phaseolus vulgaris*. Supervised by Professor Esperanza Martinez, he helped an effort trying to characterize the distribution of the symbiotic nitrogen-fixing bacteria *Rhizobium etli* inside the common bean's xylem. After finishing his coursework at UNAM, Alberto was awarded a one year fellowship to perform data analysis on high throughput sequencing of ancient DNA and RNA at the Centre for GeoGenetics, in the University of Copenhagen, Denmark. There, under the supervision of Professor Tom Gilbert, Alberto was involved in several projects aiming to describe the potential use of ancient maize and grape remains for genome-level

analyses. After finishing his internship, Alberto was admitted as a graduate student at Cornell University in the Department of Plant Breeding. There, under the supervision of Professor Edward Buckler, Alberto's research has been part of a collaborative effort lead by the International Maize and Wheat Improvement Center (CIMMYT) aiming at characterizing and guiding the efficient use of maize landraces.

A mi familia, por su apoyo incondicional. Los quiero mucho

To my husband, and to all my friends, thank you for your love, support, and for all our shared adventures

ACKNOWLEDGEMENTS

The work presented here is part of a large project called the Seeds of Discovery Initiative (SeeD). SeeD is a collaborative effort lead by the International Maize and Wheat Improvement Center (CIMMYT) in conjunction with Mexican and International partner institutions. This work was supported by SAGARPA (La Secretara de Agricultura, Ganadera, Desarrollo Rural, Pesca y Alimentacin), Mexico under the MasAgro (Sustainable Modernization of Traditional Agriculture) initiative, with additional support from the National Science Foundation (NSF), and the USDA-ARS.

I would like to acknowledge my PhD advisor Edward Buckler for giving me the opportunity to join his lab as part of the SeeD project, and for providing critical input throughout my PhD. I would also like to acknowledge Gary Atlin for his important role in organizing and designing the SeeD project, as well as helping me join the project. At CIMMYT, I would like to thank the SeeD team, lead by Sarah Hearne, Martha Wilcox on the field evaluations, Juan Burgueño on the statistical analyses of field data, and Kai Sonders on the generation of environmental data. At CIMMYT the SeeD team included also Peter Wenzl, Samuel Trachsel, Ivan Ortiz-Monasterio, Felix San Vicente, and Armando Guadarrama Espinoza, who provided important assistance in the design and evaluation of the experimental fields.

The project also included other Mexican institutions, in particular the Instituto Nacional de Investigaciones Forestales Agrcolas y Pecuarias (INIFAP), with field evaluation efforts lead by Ernesto Preciado, Arturo Terron, Humberto Vallejo Delgado, Victor Vidal, Alejandro Ortega, and Noel Orlando Gmez Montiel. At the Universidad Autonoma Agraria Antonio Narro the field trials were lead by Armando Espinoza Banda. For the work on the abnormal chromosome

10 (Ab10) I would like to thank Kelly Dawe at University of Georgia for sharing his knowledge and data on the segregation of Ab10. At University of California Davis, I would like to thank Jeff Ross-Ibarra for his suggestions on the analysis and interpretation of the data. Finally, I would like to thank at Cinta Romay and Eduardo Carrillo at Cornell University for providing input on flowering time biology and environmental ecology respectively. Also many thanks to all the members of the Buckler lab for discussion and suggestions. And very special thanks to Sara Miller for her administrative help.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Landraces, also know as farmers varieties, are domesticated populations of animal and plant species. Generally, landraces harbor within each species a significant fraction of the diversity of the domesticate. Characterizing the standing genetic variation present in landraces is critical for breeding efforts, as they provide the raw material to increase crop productivity and resilience, as well as to enhance crops' nutritional and end use value. Domestic species also offer a practical model to study the evolution of quantitative trait variation. Although collections of germplasm of landraces exist, the characterization of such diversity remains limited by its high cost, intense labor, and the difficulty to use the gained knowledge in breeding programs. This thesis focuses on the results of a comprehensive characterization of individuals from 4,500 landrace accessions representing the diversity of maize (*Zea mays* subsp. *mays*) in Latin America.

Among plant species, maize is prominent as both a globally important crop, as well as a major research model organism. As a crop, maize was domesticated around 10,000 years before present in the Balsas River in the lowlands of Mexico[1]. In the following millennia, maize was adopted by numerous cultures first in the Americas, and eventually the world[2]. Similar to the cultural and genetic heritage of maize landraces, the study of maize diversity and genetics has a very rich scientific legacy. Starting in 1876, Charles Darwin made observations of the effect of inbreeding depression for plant height in self and open pollinated maize varieties. Afterwards, in 1908, George Shull at Cold Spring Harbor made similar observations regarding inbreeding depression and the increased yield of hybrids [17], initiating the revolution of hybrid maize production in the United

States.

An important milestone for the study and collection of maize diversity in Latin America was the Mexican Agricultural Program, funded by the Rockefeller Foundation in 1943 and thoroughly documented by William Cobb in 1956[15]. As documented by Cobb, the Mexican Agricultural Program was in turn the product of decades of work from the Rockefeller Foundation in the United States and in Mexico. The pioneering projects in agriculture lead by the Rockefeller foundation trace back to 1906, when the first grant by a Rockefeller board was given to Seaman Knapp, who worked at the United States Department of Agriculture. The main goal of this grant was to improve agriculture in the United States by teaching farmers better management methods, and was motivated as a response to a highly destructive cotton pest. Since that initial grant, further efforts by the same Foundation aimed at raising the livelihoods of people through better health, and starting in 1919 those efforts included work in Mexico. In the following decades, hundreds of personnel in Mexico received public health training through the Rockefeller Foundation. Although the Rockefeller wealth was derived from profits from the petroleum industry, in Mexico the Rockefeller name was more strongly associated with the work of its Foundation.

In 1940, a new president was elected in Mexico, Manuel Camacho, representing a 20 year consolidation effort following the Mexican Revolution. Importantly, at the time the relationship between Mexico and the Unites States was tense over the recent nationalization in Mexico in 1938 of all petroleum reserves, facilities, and foreign oil companies. For the inauguration of Mexican president Camacho, the United States was represented by Vice President Henry

A. Wallace. During his visit to Mexico, Wallace represented modern scientific agriculture through the popularization of hybrid maize in the United States. The interest generated by the potential to improve maize agriculture in Mexico was evident during Wallace's stay at the US Embassy, where Mexican farmers would bring their maize ears for his advice. A year later, in 1941, Wallace met with Raymond Fosdick, President of the Rockefeller Foundation, and John A. Ferrell, director of Foundation's Public Health Activities. Wallace suggested a joint health and agriculture effort in Mexico, with the use of modern agricultural methods to increase yield of maize and beans. Fosdick appointed a group to select a team of 3 agricultural scientists to form the Survey Commission, which would then make a reconnaissance trip to Mexico. Among those in charge of choosing the Survey Commission was Albert Russell Mann, who at the time was the vice-president of the Rockefeller Foundation's General Education Board, and who had previously been the Dean of the College of Agriculture and provost at Cornell University. Mann recommended Dr. Richard Bradfield, a soil scientist, for the Survey Commission. Also recommended for the Commission was Dr. Paul C. Mangelsdorf, then a professor of economic botany at Harvard University. Mangelsdorf at the time was interested in the origin of maize, for which he had proposed a model known as the "tripartite hypothesis" where the direct ancestor of maize was extinct, and teosinte represented a hybrid between such ancestor and the sister genus *Tripsacum*. Although the model has been proven incorrect, his interest in collection of landraces for basic research was vital for the project. Finally Elvin C. Stakman was selected to join as a plant pathologist, who among other very important contributions had done significant work on the characterization of wheat black stem rust.

On July 1st, 1941 Bradfield, Mangelsdorf, and Richard E. Shultes, a recent

graduate from Harvard who had done extensive work in Mexico and was fluent in Spanish, left Ithaca, New York to Mexico in a standard green GMC suburban caryall. Stakman joined them in Mexico City on July 20th, and the party explored in total 5,000 miles of Mexico, sometimes reaching remote areas on horse or muleback. Their report to the Rockefeller Foundation was met with enthusiasm, in which the scientists made emphasis on the need for research of agricultural systems in Mexico. In 1943, the Mexican Agricultural Program was created, with J. George Harrar being appointed as the Local Director. Harrar in turn selected Edwin Wellhausen, a doctor in genetics and plant breeding from Iowa State University, to head the Mexican Agriculture Program's maize breeding program.

In the 1940s and 1950s, Wellhausen lead a major effort for collecting maize landraces across Mexico and Latin America, which became a significant portion of the Germplasm Bank of what is now the International Center for Maize and Wheat Improvement (CIMMYT). Around the same time, in 1950, Barbara McClintock published her results describing transposable elements[11], which lead to her receiving the Nobel prize in 1983. McClintock, who was part of a scientific lineage started at Cornell University by Rollins A. Emerson[10], also pioneered the use of cytological features for the inference of phylogenetic relationships among maize races and related species, and some of her subsequent analyses included the landraces collected by Wellhausen[12]. More recently, starting in the late 1980s, the landraces collected by Wellhausen and analyzed by McClintock, plus recent additional collections, were analyzed as part of the Latin American Maize Project (LAMP)[16]. The aim of LAMP was to characterize the agronomic potential harbored by this germplasm collection. This project had an important effect in the form of introgressions for disease resistance by the private sector,

lead to the establishment of the Germplasm Enhancement of Maize project in the United States, and culminated at CIMMYT with the creation of a breeder's core collection, containing the best performing landraces. The breeder's core collection, plus some additional landraces, are the samples analyzed in this thesis as part of the Seeds of Discovery(SeeD) project.

This thesis studies the evolution and adaptation of landraces through genome wide association studies. To achieve this, I first describe a novel experimental design useful for cost-effective evaluation and genotypic characterization of numerous populations. This approach is used to study flowering time and plant height. Maize wide geographic adaptation comes partially through changes in flowering time[2–6], and along with plant height, they entail important agronomic traits and have been studied recently in populations of inbred lines [6,9]. In addition, we utilize the environmental conditions from the sampling locations to study the genetic basis of local adaptation to those climatic and soil characteristics, a strategy that has been applied recently for other species[7,8]. Finally, one of the cytological markers described by Rhoades [13] and analyzed by McClintock and colleagues in maize landraces is known as abnormal 10 (Ab10). This variant of chromosome 10 displays preferential segregation during meiosis, and had a significant effect on the modern composition of the maize genome[14]. Diagnosing this feature requires observation of the chromosomes under the microscope, and here I describe a method by which the use of genetic information from populations with known Ab10 status can be used to infer the allele for individuals with only genotypic information. Although the availability of long-read genomic sequence will make the scoring of chromosomal variants easier, the method shows the potential for integrating information from different populations to infer cytological characteristics.

1.1 References

1. Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J. & Dickau, R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas.

2. Mir, C. et al. Out of America: tracing the genetic footprints of the global diffusion of maize. Theor. Appl. Genet. 126, 26712682 (2013).

3. Bertin, P., Madur, D., Combes, V., Dumas, F. & Brunel, D. Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. PLoS One (2013).

4. Ducrocq, S. et al. Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178, 24332437 (2008).

5. Hung, H.-Y. et al. ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc. Natl. Acad. Sci. U. S. A. 109, E191321 (2012).

6. Buckler, E. S. et al. The genetic architecture of maize flowering time. Science 325, 714718 (2009).

7. Fournier-Level, A. et al. A map of local adaptation in Arabidopsis thaliana. Science 334, 8689 (2011).

8. Lasky, J. R. et al. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv 1, e1400218 (2015).

9. Peiffer, J. A. et al. The genetic architecture of maize height. Genetics 196, 13371356 (2014).

10. Dove, W. F. Anecdotal, Historical and Critical Commentaries on Genetics.

Genetics 135, 937 (1993).

11. McClintock, B. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences 36, 344355 (1950).

12. McClintock, B., Yamakake, T. A. K. & Blumenschein, A. Chromosome constitution of races of maize: its significance in the interpretation of relationships between races and varieties in the Americas. (Colegio de Postgraduados Mexico, 1981).

13. Rhoades, M. M. Preferential Segregation in Maize. Genetics 27, 395407 (1942).

14. Buckler, E. S., 4th et al. Meiotic drive of chromosomal knobs reshaped the maize genome. Genetics 153, 415426 (1999).

15. William C. Cobb, The historical background of the Mexican Agricultural Program, 1956

16. Salhuana, W., Jones, Q. & Sevilla, R. The Latin American Maize Project: Model for rescue and use of irreplaceable germplasm. Diversity (1991).

17. Crow, J. F. 90 years ago: the beginning of hybrid maize. Genetics 148, 923928 (1998).

CHAPTER 2

IDENTIFYING THE DIAMOND IN THE ROUGH: A STUDY OF ALLELIC DIVERSITY UNDERLYING FLOWERING TIME ADAPTATION IN MAIZE LANDRACES

2.1 Abstract

Landraces (traditional varieties) of crop species are a reservoir of useful genetic diversity, yet often remain untapped due to the genetic linkage between the few useful alleles with hundreds of undesirable alleles[1]. We integrated two approaches to characterize the genetic diversity of over 3000 maize landraces from across the Americas. First, we mapped the genomic regions controlling latitudinal and altitudinal adaptation, identifying 1,498 genes. Second, we developed and used F-One Association Mapping (FOAM) to directly map genes controlling flowering time across 22 environments, identifying 1,005 genes. In total 65% of the SNPs associated with altitude were also associated with flowering time. In particular, we observed many of the significant SNPs were contained in large structural variants (inversions, centromeres, and pericentromeric regions): 29.4% for flowering time, 58.4% for altitude and 13.1% for latitude. The combined mapping results indicate that while floral regulatory network genes contribute substantially to field variation, over 90% of contributing genes likely have indirect effects. Our strategy can be used to harness the diversity of maize and other plant and animal species.

[0]For this chapter, my contribution encompassed analyses using the genotypic data, as well as the genome wide association models and subsequent gene level analyses

2.2 Results

Maize (*Zea mays* subsp. *mays*) is a model organism with a legacy of a hundred years of cytological, genetic, and biomolecular characterization[2]. Maize displays high levels of genetic diversity with low linkage disequilibrium (LD)[3,4], low population differentiation[5], prevalent migration[6] and occasional introgression from wild relatives[7–9]. More recently, experimental populations, like the Nested Association Mapping (NAM) populations[10,11], and large association panels[4,12] have allowed mapping and deployment of useful alleles for several quantitative traits[13–16]. However, most of the founder lines from these panels correspond to highly inbred improved lines, many from temperate regions, capturing only a modest fraction of the total diversity present in the species. In contrast, maize landraces span numerous ecogeographic areas and harbor most of the diversity of the species. Nevertheless, maize landraces, like many other crops' traditional varieties, remain largely uncharacterized by genomics.

This study maps genes controlling flowering time with two distinct methods: (1) Each of these landraces come from environments to which they are well adapted. We used this adaptation as the trait to identify genes driving large scale adaptation. (2) We mapped flowering time variation in controlled field experiments through a novel, rapid, experimental design called F-One Association Mapping (FOAM) (Figure 2.1). Briefly, FOAM consists of sampling single individuals across numerous populations, which are genotyped and crossed to one or a small number of common parents to derive F1 families. Subsequently GWAS is performed from multi-trial F1 progeny evaluation. Major advantages for this design are (a) capturing thousands of alleles across populations, (b) maintaining the tractability of two alleles per loci per individual, and (c) am-

ple replication of alleles increasing the power and accuracy for genetic effect estimation. The main limitation of FOAM is that the nested evaluation of different subsets of F1 progeny by ecological zone limit the ability to accurately estimate genotype by environment interaction effects.

Our maize landrace FOAM population used individuals from 4,471 accessions from 35 countries in the Americas (Figure 2.2) grouped into three adaptation classes to account for altitude adaptation (low, middle and high elevation). Similarly, the common parents and evaluation sites were nested within adaptation class (Methods, (Figure 2.6)[17,18]. Landrace parents were genotyped for close to one million SNPs using Genotyping by Sequencing[19], and missing data was imputed using BEAGLE4[20]. Of the 4,471 accessions, 3,552 yielded F1 families containing both genotypic profiles and sufficient progeny, 3,633 contained detailed passport information which was used for mapping large scale adaptation, and 2,603 were present in both mapping studies.

We first explored the effects of recombination frequency and geography-driven limited dispersal on the distribution of genetic diversity in the landrace parents. Using Multidimensional Scaling (MDS, Methods), we observed the first axis and second axis explained only 6.1% and 1.7% of of the variance respectively, consistent with the low F_{ST} in maize landraces[5]. The first axis separates among Mexican landraces, consistent with Mexican landraces having a deeper coalescent and greater representation in the panel. The second axis was associated to a latitudinal North to South gradient across Latin America representing isolation by distance (Figure 2.7). In addition, a Mantel test[21] revealed a significant correlation between geographic and genetic distances (Pearson's r= 0.46, p-value<0.001), with most of the association driven by altitude. De-

spite this, phylogenetic analysis (Methods, (Figure 2.8)) revealed that adaptation class does not drive clade membership, which indicates that alleles segregate across adaptation classes, with highland adaption being polyphyletic, consistent with recent reports[22]. To study recombination, we estimated an approximate LD statistic (Methods) which shows a distribution consistent with previous recombination estimates[23,24], with higher recombination in gene-rich regions, and lower around centromeres. Each chromosome displayed a unique recombination landscape, with the presence of half a dozen high LD regions (Figure 2.9), which together encompassed 6.1% of the base pairs of genome, accounting for 2.8% of the annotated coding genes. Together, these results suggest that although at large scale geography and adaptation contribute to the distribution of diversity, even with the large effective population size of landraces at the genomic scale a complex recombination landscape limits the free segregation of alleles through increased LD.

Flowering time generally plays a crucial role in local adaptation of plants, and in maize flowering time is a complex trait controlled by hundreds of small effect loci, many with rich allelic series [4,14,25–30]. We used altitude and latitude from sampling location as traits for mapping local adaptation, and the significance thresholds were chosen to maximize overlap rate between flowering, altitude, and latitude genes (Methods, (Figure 2.10). For altitude, we observed 58.4% of the significant SNPs corresponded to regions with higher LD. In particular, INV4m, the 13Mb adaptive introgression from highland teosinte into maize[8,31], was highly significant. We also observed significance for the centromeres of chromosomes 2,5,6,8 and a large region upstream of the centromere on chromosome 3. Outside this low recombination regions, 366 genes showed significant association with altitude. For latitude, we observed 13.1% of the

significant markers were contained within low recombination regions, particularly the centromere of chromosome 5. In total across all Latin America, 1,498 genes showed significant association with latitude, of which 395 of were shared with altitude. The minor allele frequency distribution of the significant alleles indicated that many are shared across clades and landraces, which was very distinct from the neutral distribution (Figure 2.3). These 1,498 genes appear to be the main contributor to large scale environmental adaptation to altitude and latitude - the key drivers of flowering time.

To study the genetic basis of flowering time, we conducted field evaluation on F1 progeny across 22 trials and 2 years in 13 locations across Mexico, with each trial containing a different subset of the collection to maximize number of accessions evaluated (Methods, (Figure 2.1). Phenotypic data was analyzed independently for each trial using a mixed linear model (Methods, equation 2.1), yielding 18,797 accession parent-environment estimates for each male and female flowering time. We performed genome wide association for days to male and female flowering using a mixed linear model (Methods). In total 72% of the associated SNPs were significant for both male and female flowering, as expected from the overlapping genetic control[14]. There was a significant contribution of low recombination regions in flowering time variation, parallel to that of latitude and longitude, with a 20-fold enrichment for significant SNPs at high LD regions (Pearson's chi-squared, p-value <2.2e-16). In particular, significant variants included the centromeres of chromosomes 3, 5, and 6, INV4m, and a 6Mb region on chromosome 3 beginning at 79Mb. The 6Mb region on chromosome 3 has a segregation similar to INV4, and together with its increased LD suggests it might be an inversion. In NAM this putative inversion and the centromere comprise a single QTL for flowering time[14]. For the centromere of

chromosome 5 there were 3 distinctive alleles segregating in the landraces, all present in the NAM population (Figure 2.11). The inverted allele of INV4m, although absent in temperate material, segregates at high frequency in highland landraces (Figure 2.12), where it has a very large additive effect advancing flowering by three days, the largest effect for flowering time in maize to date. Both homozygous alleles from the putative inversion on chromosome 3 segregate across our maize landrace panel and the NAM population. Compared to INV4m, this locus displays a more modest effect on flowering time. The heterotic effect of the centromere of chromosome 5 on yield[32],potentially product the complementation of deleterious mutations[23], suggests that the significant inversions and centromeres may similarly affect flowering time through heterotic effects leading to more vigorous plants, which in maize generally results in earlier flowering.

Outside the structural variants, we observed 881 and 883 genes (around 2.2% of genes) with significant association for days to female and days to male flowering respectively (Figure 2.4). To further characterize the regions associated with flowering time, we looked for gene ontology enrichment and gene expression using the maize transcription atlas[33](Methods), and compared the significant genes to a candidate gene list containing genes characterized in other populations, known to interact in the maize flowering time regulatory network[34] as well as the 25 members of the *Zea mays*2 CENTRORADIALIS (ZCN) gene family[35]. Overall the associating genes tended to be expressed in anthers, and enriched in general metabolic processes, with the genes known to be part of the regulatory network being more expressed in immature cob and the tip of the leaf at V5 stage and enriched for regulatory processes (Figure 2.13). We observed a significant enrichment in flowering time candidate genes compared to the rest

of the genome (Fisher's Exact Test p-value = 4.3 x 10-7). In total 10 and 12 candidate genes representing the circadian clock, photoperiod, gibberellin acid, and circadian clock pathways displayed significant associations with male and female flowering respectively. Out of these, nine were common for both types of flowering. The most significant hits corresponded to *Vgt1*[36,37], one of the largest known GE QTL, and *Zcn8*[35,38], the maize florigen and homolog to *FT* in *Arabidopsis* (Figure 2.5). *ZmCCT*, the largest photoperiod QTL[29], was only modestly significant for latitude, and significant only for days to female flowering, probably a result of non-inducing sampling and trial locations. In particular for the gene *d8*, a locus with cryptic association with flowering time[34], we observed significance for this gene around 50 and 100kb up and downstream the coding region for latitude, altitude, and both male and female flowering, overlapping with the region previously observed to display divergent selection associated with climate adaptation[39]. In addition, the distribution of the flowering time associating genes displayed a significant geography effect, with 56 and 52 genes in common with altitude and latitude respectively. In general, the minor alleles for flowering time tended to be associated with high elevation, and northwest coordinates, however the minor allele frequency distribution of the significant SNPs was different to that of the alleles significant for altitude and latitude, having a significant shift towards low frequency polymorphisms (Figure 2.3). Together, these results support the model of infrequent variants in recurrent regulatory genes underlying the genetic control of flowering time variation in maize, with adaptive alleles segregating across populations, and their distribution matching the fitness optimum according to geographic variation. In particular, the high overlap between significant SNPs for altitude and flowering time suggests that for tropical maize flowering time adaptation is very relevant for changes

in elevation, which affects among others spectral composition and intensity of incident light, as well as the incidence of heat and cold stress. In contrast, the lower overlap between latitudinal and flowering time associating SNPs could be due to the sampling from non-photoperiod-inducing latitudes, potentially leading to latitudinal flowering time adaptation being relevant for other biotic (disease pressure) and abiotic (soil pH, precipitation) stresses.

We assayed the potential for predicting flowering time in the landraces using either all of the high density genetic markers or just the markers significantly associated with the trait. We performed genome wide prediction using gBLUP independently for each trial (Methods). The average 5-fold cross-validated prediction accuracy was 0.45 across trials for both male and female flowering time, and as high as 0.7 for some trials (Figure 2.14). Genomic prediction accuracy between the top genes from GWAS was equivalent to that of 30,000 random evenly distributed SNPs, highlighting their potential use for breeding of the significant markers. Intriguingly prediction accuracy was not correlated with our other heritability estimate (Pearson cor=0.22), which could be an effect of the differences in the genetic variances and sample sizes across all trials. Together the good predictive ability of the significant regions for genomic selection shows the potential to greatly speed the breeding of new adapted varieties with exotic beneficial alleles.

Crop landraces are an incredible source of diversity that will be necessary to adapt our crops to the next century of climate change. However, their tremendous diversity and genetic load prevent them from being efficiently tapped without a genomic index. This research lays out two complementary strategies for tapping this diversity. The geographic associations have high statistical

power for identifying the adaptive loci, which appear to be common and shared, and are unlikely to be deleterious given their high frequency. This extensive sharing is probably the result of outcrossing and extensive migration throughout Latin America in last several millennia. The limitation of this approach is that correlated traits and adaptations are being co-mapped. The novel FOAM field trial associations, while substantially overlapping, are showing the impacts of deleterious and private mutations and their complementation in these hybrid trials. These deleterious alleles have been the bane of breeders wanting to tap landrace diversity. The strategy for tapping this diversity should use the overlapping genes and alleles of the two separate approaches, as these have proven to be adaptive and target the trait of interest. In addition, breeding efforts could incorporate standard genomic selection coupled with genome editing. This provides an efficient strategy to tap landraces' diversity and allow our crops to adapt to faster changes than they have ever had in the past.

2.3 Methods

Mating design and phenotypic evaluation

The mating design for the maize landrace FOAM population consisted of crossing each accession male to single cross hybrid females of matching adaptation. Leaf tissue of the landrace individual was collected for genotyping. The progeny evaluation trials were performed across 2 years in 13 locations across Mexico using an augmented row-column design, which includes systematic checks in field rows and columns[40]. There were between 288 and 1,928 accessions per trial, with an average of 834 (Figure 2.1). Over half of the accessions

were replicated in 5 or more trials, with a maximum value of 13 trials per accession and a minimum of 1. For each trial, each experimental row contained between 10 and 25 progeny plants. The replication across trials together with the use of systematic checks across experimental fields provides sufficient allelic replication for accurate estimation of genetic effects. Flowering time was measured in each trial following the maize standard, i.e. the number of days from planting until half of the individuals within a plot displayed silks for female flowering or anthers in half of the central spike for male flowering.

Genotyping

Accessions used as male parents were genotyped using GBS[19], with ApeKI as the restriction enzyme to a replication level of 96 individuals per sequencing plate. Approximately $8x10^9$ sequencing reads were generated using an Illumina HiSeq for the landrace accessions and sequence reads were analyzed jointly with another 40,000 maize lines as part of the GBS Build 2.7 using TASSEL[41]. For association analyses, missing data was imputed using BEAGLE420, which has been shown to yield the best current accuracies in maize heterozygous material (R^2=0.68)[42]. After imputation, SNPs were filtered for minor allele frequency greater than 1% resulting in approximately 500,000 biallelic markers across the genome. GBS non-imputed markers can be accessed at http://hdl.handle.net/11529/10034 and imputed GBS markers at http://hdl.handle.net/11529/10035

Diversity Assessment

For the Mantel test[21], we calculated the pairwise Euclidean distance matrix based on the geographical data from the accessions (latitude, longitude, and altitude, http://mgb.cimmyt.org/gringlobal/search.aspx). The genetic distance matrix was estimated from a genome wide random sample 30,000 non-imputed markers using TASSEL. The distance matrix was used for estimating the Neighbor-Joining tree using TASSEL. Multidimensional Scaling (MDS) was performed on the genetic distance matrix using the cmds function in R.

Recombination

Our LD statistic consisted in estimating the correlation between markers across the genome at 100 site windows using all homozygote and heterozygote non-imputed markers with the LD function on the software TASSEL. For comparing the LD and recombination values, we estimated the correlation at 1Mb sliding windows between (1) the log10 median LD estimate (2) the log median crossover probabilities estimated using the American and Chinese Nested Association Mapping populations[23], and (3) the log median population recombination rates (rho) estimated both for improved lines and landraces Hapmap2 project[24]. Our LD estimates displayed a negative correlation with gene density (r=-0.57) and NAM crossover probability23 (r=-0.45). We observed a modest negative correlation (r=-0.33) with a population genetic estimate of historical recombination (rho) [23,24]. High-LD regions were defined based on the change in slope of global median LD (Figure 2.9) as those segments that had a median LD greater than 0.01. In total, there were 256 high LD regions encompassing 7.8% of the genome.

Genome wide association with altitude and latitude

We performed Genome wide association using a generalized linear model with altitude and latitude as response variables and markers filtered at 1% frequency as explanatory variables. Altitude and latitude were recorded during field sampling of the original accessions. In order to establish a significance threshold to avoid excess of false positives, we estimated the overlap rate using the most significant flowering time GWAS SNPs. Overlap Rate was defined as the set of overlapping SNPs between the top flowering time SNPs and either altitude or latitude, divided by the union of the sets across significance thresholds from 0.001 to 0.01. Significance thresholds chosen were 0.005 for altitude and 0.01 for latitude ((Figure 2.10). Heritability estimates were 0.88 for altitude and 0.85 for latitude, estimated LDAK[43] with a single Kinship matrix, estimated with all the Beagle4 imputed markers, and the matrix was estimated from the algorithm implemented in GCTA[44].

Analyses of structural variants

In order to infer the underlying haplotypes for the centromeres of chromosomes 3,5,6, as well as INV4 and the high-LD region on chromosome 3, we first estimated a genetic distance matrix for each locus using the non-imputed markers. The distance matrices were then analysed using multidimensional scaling. The centromere of chromosome 5 segregates in the landraces with three distinct homozygous haplotypes and their corresponding heterozygote pairs. The region around the centromere of chromosome 6 was 12 Mb in size, includes the centromere and a large pericentromeric region that expands out in both directions; it displayed a similar pattern to the centromere of chromosome 5, however

distinct alleles were not called due to the excess of heterozygous individuals between the homozygous classes, probably reflecting recombinant haplotypes. The centromere of chromosome 3 displayed a more complex pattern of distance than the other two associating centromeres, likely due to the presence of more than three segregating haplotypes. For INV4, we observe two distinct alleles and the heterozygote. We observed the allele is fixed in many of CIMMYT's improved lines (Table 3), including those used as parents for the highland test crosses in the present experiment.

Analysis of phenotypic data

To estimate the breeding values of the landrace accession parent, for each trial a mixed linear model was fitted using restricted maximum likelihood method, in ASREML (V 3.0), using the progeny's calendar days to male or female flowering as a response variable. Of the 23 trials planted, one was excluded because flowering time data was not collected according to protocol. The models included fixed effects for checks, tester, and hybrid and a random effect of accession in a complete nested model. Including in the model the random effect of row and column and using an autoregressive model of order 1 in row and columns controlled experimental noise product of field variation. All the random effects were considered independent one from each other. The model used can be expressed as follows:

$$y_{ijklm} = \mu + \gamma_i + \lambda_j + \alpha_k + \beta_{l(k)} + \delta_{m(kl)} + \epsilon_{ij} \tag{2.1}$$

where y_{ijklm} is the phenotype, μ is the overall mean, γ_i is the effect of the i-th row $\gamma_i \sim N(0, \sigma_1^2)$, λ_j is the effect of the j-th row $\lambda_j \sim N(0, \sigma_2^2)$, α_k is the effect of the

k-th group, k=1,...,K, K+1, if k ≤ K the group is a check, the group K+1 is the average of the testers, $\beta_{l(k)}$ is the effect of the l-th tester in group K+1, $\delta_{m(kl)}$ is the effect of the m-th accession in the tester k in group K+1, $\delta_{m(kl)} \sim N(0, \sigma^2_{kl})$, and ϵ_{ij} is the experimental error.

For the experimental error we assume the following distribution:

$\epsilon \sim N(0, \Sigma)$, with $\Sigma = \Sigma_r \otimes \Sigma_c$ and

$$\Sigma_r = \begin{bmatrix} 1 & \rho_r^1 & \rho_r^2 & \dots & \rho_r^{d-2} & \rho_r^{d-1} \\ \rho_r^1 & 1 & \rho_r^1 & \dots & \rho_r^{d-3} & \rho_r^{d-2} \\ \rho_r^2 & \rho_r^1 & 1 & \dots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_r^{d-2} & \rho_r^{d-3} & \rho_r^{d-4} & \dots & 1 & \rho_r^1 \\ \rho_r^{d-1} & \rho_r^{d-2} & \rho_r^{d-3} & \dots & \rho_r^1 & 1 \end{bmatrix} \Sigma_c = \begin{bmatrix} 1 & \rho_c^1 & \rho_c^2 & \dots & \rho_c^{d-2} & \rho_c^{d-1} \\ \rho_c^1 & 1 & \rho_c^1 & \dots & \rho_c^{d-3} & \rho_c^{d-2} \\ \rho_c^2 & \rho_c^1 & 1 & \dots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_c^{d-2} & \rho_c^{d-3} & \rho_c^{d-4} & \dots & 1 & \rho_c^1 \\ \rho_c^{d-1} & \rho_c^{d-2} & \rho_c^{d-3} & \dots & \rho_c^1 & 1 \end{bmatrix}$$

Genome wide association with flowering time

Association analysis was performed in two steps for all trials using a linear mixed model. For each trait (days to male and female flowering) two models were fitted, one with the trait BLUPs as response variable and another with the standardized values of the same BLUPs. This was done in the absence of growing degree units, to verify the consistency of the results given the uneven variances for the trait across the various trials. The first step models included the fixed effects for trial (categorical); population structure in the form of 10 MDS weights (numerical) that together explained around 13% of the genetic and 10.6% of the phenotypic variances; and the effect of the hybrid used as parent for each accession's cross. The random effect of related-

ness was added to both models in the form of a kinship matrix. The kinship matrix was estimated using the same subset of SNPs as the MDS weights. The mixed model was fit using the R package EMMREML (http://cran.r-project.org/web/packages/EMMREML/index.html). Residuals were obtained from those models and fitted in the second step models as a response variable for the single marker analysis using R, with marker nested within trial. The model equation used was:

$$y_{ijk} = \mu + T_i + H_j + Q_{ijk} + K + \epsilon_{ijk} \tag{2.2}$$

where y_{ijk} are the breeding values for flowering time, μ is the overall mean, T_i is the main effect for trial, H_j is the main effect for hybrid parent of testcross individuals, Q_{ijk} are 10 MDS weights controlling for population structure, K is the kinship matrix and ϵ_{ijk} is the random experimental error

In the second step of the association model, the residuals from the first model were fitted as a response variable in the model

$$Y = S_{[t]} + \epsilon_i \tag{2.3}$$

Where Y is the residual, S is the SNP effect and is nested within trial t. The model tests the null hypothesis $H_0 : S = 0$ that the effect of each SNP is 0 in all trials.The alternative hypothesis is that the SNP has an effect on any trial. The reason for testing this hypothesis is that the effect of each SNP can, and often does, change on value and direction depending on its segregation on the population and its phase with the causal polymorphism. We consider as significant the top one percent of the SNPs based on p-value, which all had -log10 p-values

greater than 18.

Significance at genic regions

We reasoned that significance at candidate genes would depend on local LD and genotype coverage, therefore a higher proportion of significant SNPs around candidate genes would be indicative of association at the gene itself rather than at the entire LD block or because of higher genotype coverage. On that account, we looked at significant associating SNPs within 50 kb up and downstream of candidate genes. Of all the candidate genes, only PhyB1, GL15 and ZCN13 are in the high-LD set and therefore were excluded from this analysis. Genome wide prediction was performed with using the software GAPIT[43]. The models were run for each trial, and accuracy was measured from performing 5 fold cross validation in 10 replicates for each trial. Two models were run for each trait/trial. One model used a kinship matrix estimated 1 SNP for each of the associated genomic regions, while the other used 30,000 random SNPs for the estimation of the kinship matrix. All models included 10 MDS weights to account for population structure.

Expression across tissues

We used the transcription data from the maize atlas[33] for the following 11 tissues: 16 days after pollination embryo, 16 days after pollination endosperm, 6 days after silking primary root, tip of stage 2 leaf at V5 plant stage, base of stage 2 leaf at V5 plant stage, 13th leaf at V9 stage, 13th leaf at R2 stage , silk, anthers, Immature cob at V18 stage, 4th internode at V9 stage, and stem and shoot apical

meristem at V4 stage. We used the standardized expression values, and estimated for each gene what tissue it was most expressed at. We then performed a chi-squared test for each tissue to test if there were more genes expressed at the candidate or associating genes than expected under the null model of equal levels of the global expression pattern.

2.4 References

1. Warburton, M. L. et al. Genetic Diversity in CIMMYT Nontemperate Maize Germplasm: Landraces, Open Pollinated Varieties, and Inbred Lines. Crop Sci. 48, 617 (2008).

2. Wallace, J. G., Larsson, S. J. & Buckler, E. S. Entering the second century of maize quantitative genetics. Heredity 112, 30-38 (2014).

3. Remington, D. L. et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. U. S. A. 98, 11479-11484 (2001).

4. Romay, M. C. et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14, R55 (2013).

5. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. Nat. Genet. 44, 808-811 (2012).

6. Mir, C. et al. Out of America: tracing the genetic footprints of the global diffusion of maize. Theor. Appl. Genet. 126, 2671-2682 (2013).

7. van Heerwaarden, J. et al. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc. Natl. Acad. Sci. U. S. A. 108, 1088-1092 (2011).

8. Hufford, M. B. et al. The genomic signature of crop-wild introgression in

maize. PLoS Genet. 9, e1003477 (2013).

9. Warburton, M. L. et al. Gene flow among different teosinte taxa and into the domesticated maize gene pool. Genet. Resour. Crop Evol. 58, 1243-1261 (2011).

10. McMullen, M. D. et al. Genetic properties of the maize nested association mapping population. Science 325, 737-740 (2009).

11. Li, C. et al. Quantitative trait loci mapping for yield components and kernel-related traits in multiple connected RIL populations in maize. Euphytica 193, 303-316 (2013).

12. Flint-Garcia, S. A. et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J. 44, 1054-1064 (2005).

13. Peiffer, J. A. et al. The genetic architecture of maize height. Genetics 196, 1337-1356 (2014).

14. Buckler, E. S. et al. The genetic architecture of maize flowering time. Science 325, 714-718 (2009).

15. Harjes, C. E. et al. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319, 330-333 (2008).

16. Tian, F. et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43, 159-162 (2011).

17. Salhuana, W., Jones, Q. & Sevilla, R. The Latin American Maize Project: Model for rescue and use of irreplaceable germplasm. Diversity (1991).

18. Pollak, L. M. The history and success of the public-private project on germplasm enhancement of maize (GEM). Adv. Agron. (2003).

19. Elshire, R. J. et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One 6, e19379 (2011).

20. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459-471 (2013).

21. Mantel, N. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209-220 (1967).

22. Takuno, S. et al. Independent molecular basis of convergent highland adaptation in maize. bioRxiv (2015). doi:10.1101/013607

23. Rodgers-Melnick, E. et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. U. S. A. 112, 3823-3828 (2015).

24. Chia, J.-M. et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803-807 (2012).

25. Bertin, P., Madur, D., Combes, V., Dumas, F. & Brunel, D. Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus. PLoS One (2013).

26. Ducrocq, S. et al. Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178, 2433-2437 (2008).

27. Hirsch, C. N. et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121-135 (2014).

28. Chardon, F. et al. Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. Genetics 168, 2169-2185 (2004).

29. Hung, H.-Y. et al. ZmCCT and the genetic basis of day-length adaptation underlying the postdomestication spread of maize. Proc. Natl. Acad. Sci. U. S. A. 109, E1913-21 (2012).

30. Salvi, S. et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. U. S. A.

104, 11376-11381 (2007).

31. Pyhjrvi, T., Hufford, M. B., Mezmouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. Genome Biol. Evol. 5, 1594-1609 (2013).

& 32. Stuber, C. W., Lincoln, S. E., Wolff, D. W., Helentjaris, T. & Lander, E. S. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132, 823-839 (1992).

33. Sekhon, R. S. et al. Genome-wide atlas of transcription during maize development. Plant J. 66, 553-563 (2011).

34. Dong, Z. et al. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One 7, e43450 (2012).

35. Danilevskaya, O. N., Meng, X., Hou, Z., Ananiev, E. V. & Simmons, C. R. A genomic and expression compendium of the expanded PEBP gene family from maize. Plant Physiol. 146, 250-264 (2008).

36. Salvi, S. et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. U. S. A. 104, 11376-11381 (2007).

37. Castelletti, S., Tuberosa, R., Pindo, M. & Salvi, S. A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. G3 4, 805-812 (2014).

38. Meng, X., Muszynski, M. G. & Danilevskaya, O. N. The FT-like ZCN8 Gene Functions as a Floral Activator and Is Involved in Photoperiod Sensitivity in Maize. Plant Cell 23, 942-960 (2011).

39. Camus-Kulandaivelu, L. et al. Patterns of molecular evolution associated with two selective sweeps in the Tb1-Dwarf8 region in maize. Genetics 180, 1107-1121 (2008).

40. Crossa, J. & Federer, W. T. I.4 Screening Experimental Designs for Quanti-

tative Trait Loci, Association Mapping, Genotype-by Environment Interaction, and Other Investigations. Front. Physiol. 3, (2012).

41. Glaubitz, J. C. et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One 9, e90346 (2014).

42. Swarts, K., Li, H., Romero Navarro, J. A. & An, D. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant 7, (2014).

43. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 24, 1550-1557 (2014).

44. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76-82 (2011).

45. Lipka, A. E. et al. GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397-2399 (2012).

2.5 Figures

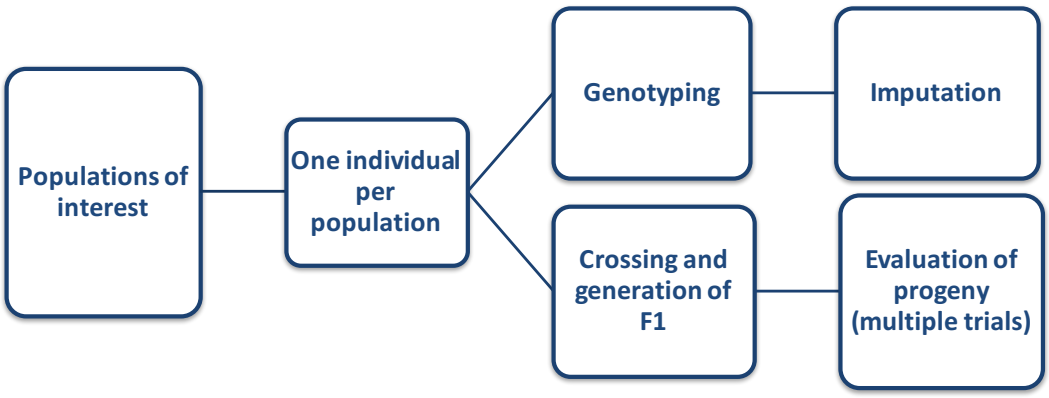


Figure 2.1: FOAM experimental design

2.1 General experimental design for F-One Association Mapping populations. One individual from each of up to thousands of individuals is genotyped and used as parent. Progeny are then evaluated for multiple years/locations to estimate the genetic contribution of the original individual and phenotypic and genotypic data are used for Genome Wide Association

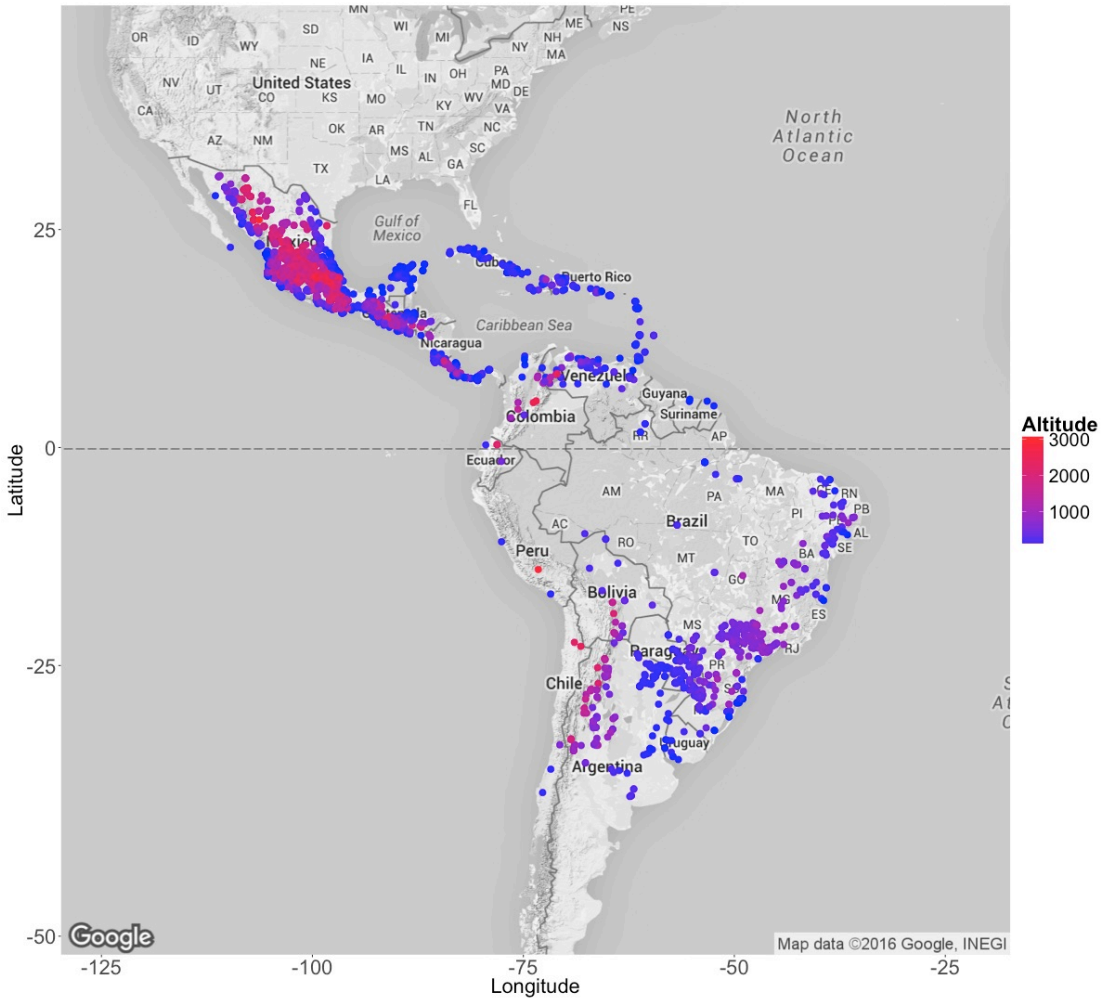


Figure 2.2: Sampling location of landrace accessions

2.2 Geographic coordinates of original sampling sites of landrace accessions across all Latin America. Color gradient corresponds to altitude.

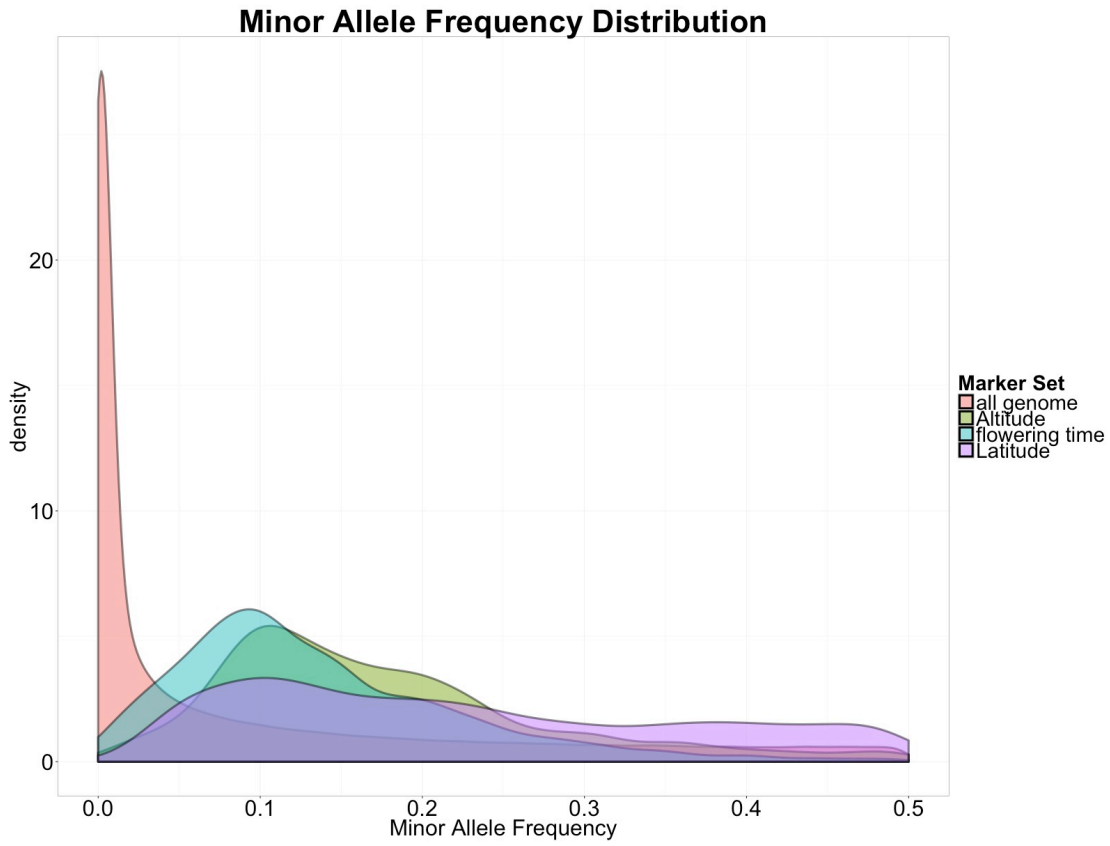
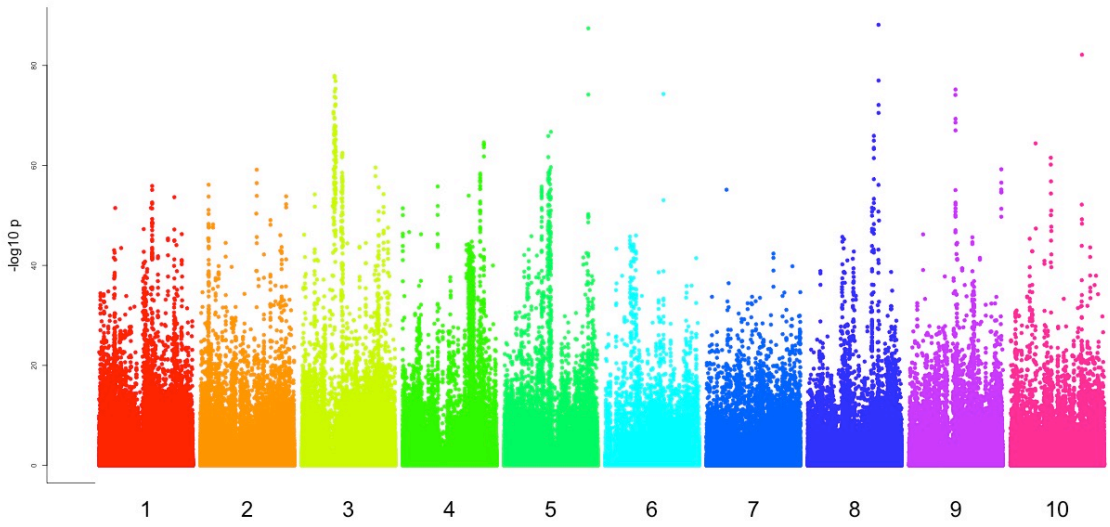


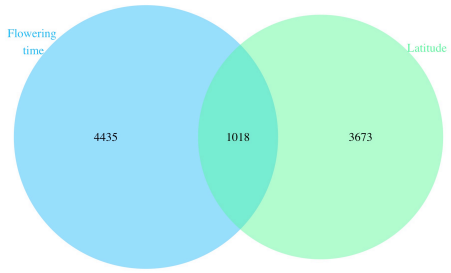
Figure 2.3: Genomewide minor allele frequency distributions

2.3 Minor Allele frequency distribution for all segregating SNPs, as well as the most significant SNPs for each trait. The minor allele frequency of the genome functions as a null distribution, and is enriched at low frequency polymorphisms as expected for a population under selection-mutation balance. The SNPs associated with Latitude show the most significant deviation from the null, with most variation at high frequency. Altitude associating SNPs are also enriched at high frequency, with most of the density between MAF of 0.1 and 0.2. Finally, flowering time associating SNPs represent a mix between low and high frequency polymorphisms, probably reflecting deleterious and adaptive alleles respectively.

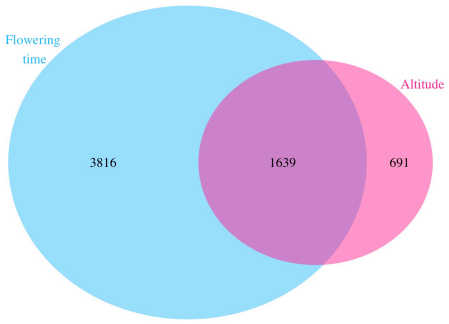
2.4 a) Manhattan plot for Days to anthesis, with the ten chromosome of maize on the x-axis and the -log10 p value on the y axis. The most significant hits on chromosome 8 correspond to VGT1 and ZCN8, and the significant peak on chromosome 3 corresponds to a previously unreported inversion. b) Overlap between significant SNPs for flowering time and latidue and c) altitude, with SNPs associated with altitude having significantly more overlap with flowering time associated SNPs d) Local Manhattan plot for chromosome 4 for days to silking and e) altitude. The large region with significance corresponds to INV4m, the adaptive introgression from highland teosinte to highland maize



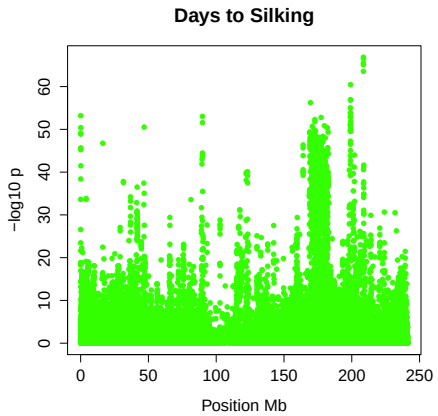
(a) Manhattan plot for days to anthesis



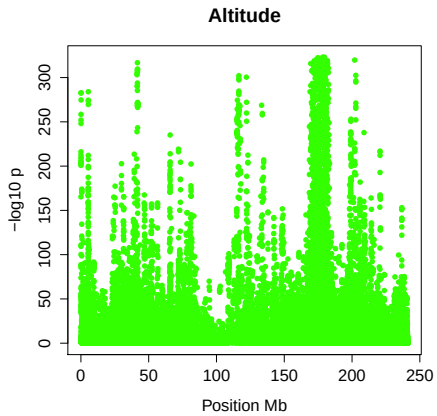
(b) Overlap flowering time latitude



(c) Overlap flowering time altitude



(d) Chromosome 4 Days to silking



(e) Chromosome 4 Altitude

Figure 2.4: Manhattan plots and overlap between traits

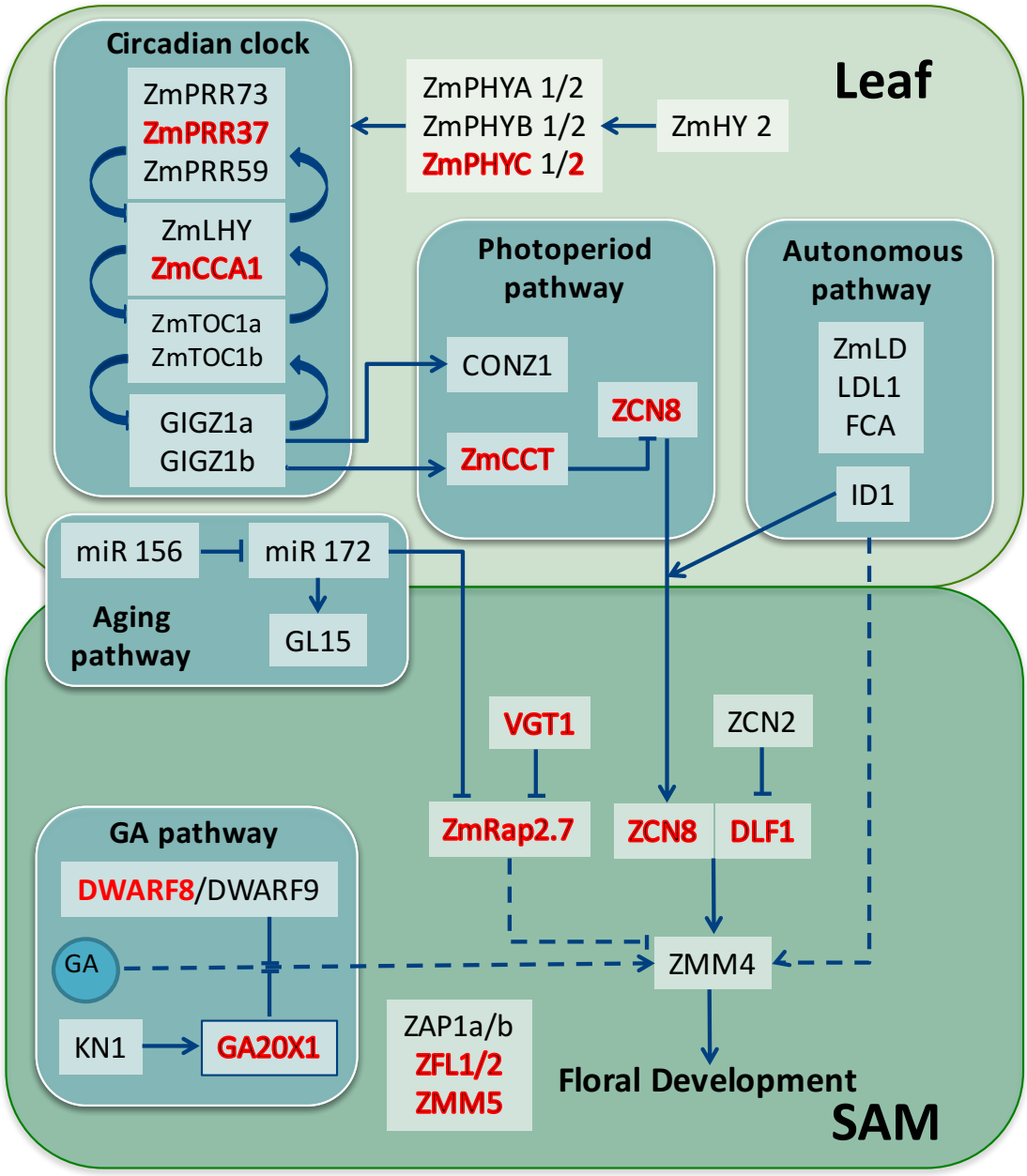


Figure 2.5: Significant genes within the flowering time pathway

2.5 Flowering time pathway, modified from Dong, et al[34], showing the genes involved in flowering time at the leaf and Shoot Apical Meristem (SAM). The genes highlighted in red displayed significant association with flowering time in our study

2.6 Supplemental material

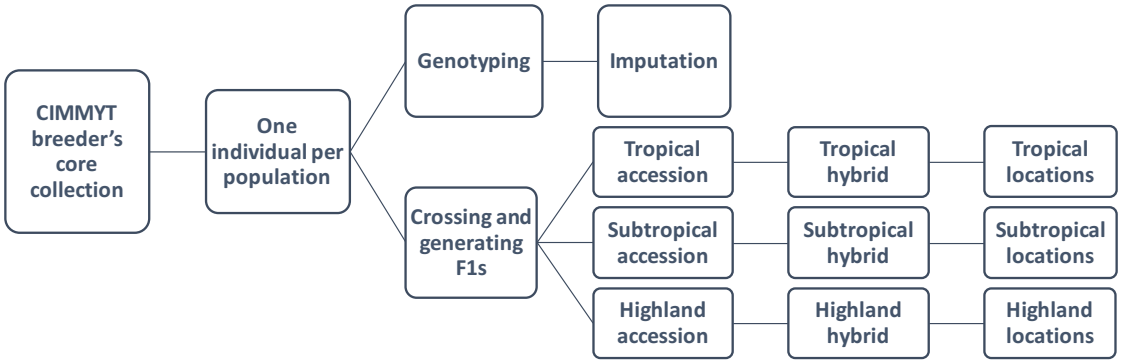


Figure 2.6: FOAM design nested within adaptation

2.6 Maize custom landrace FOAM design. Because of the large effect of adaptation, crossing and evaluation are nested within adaptation classes

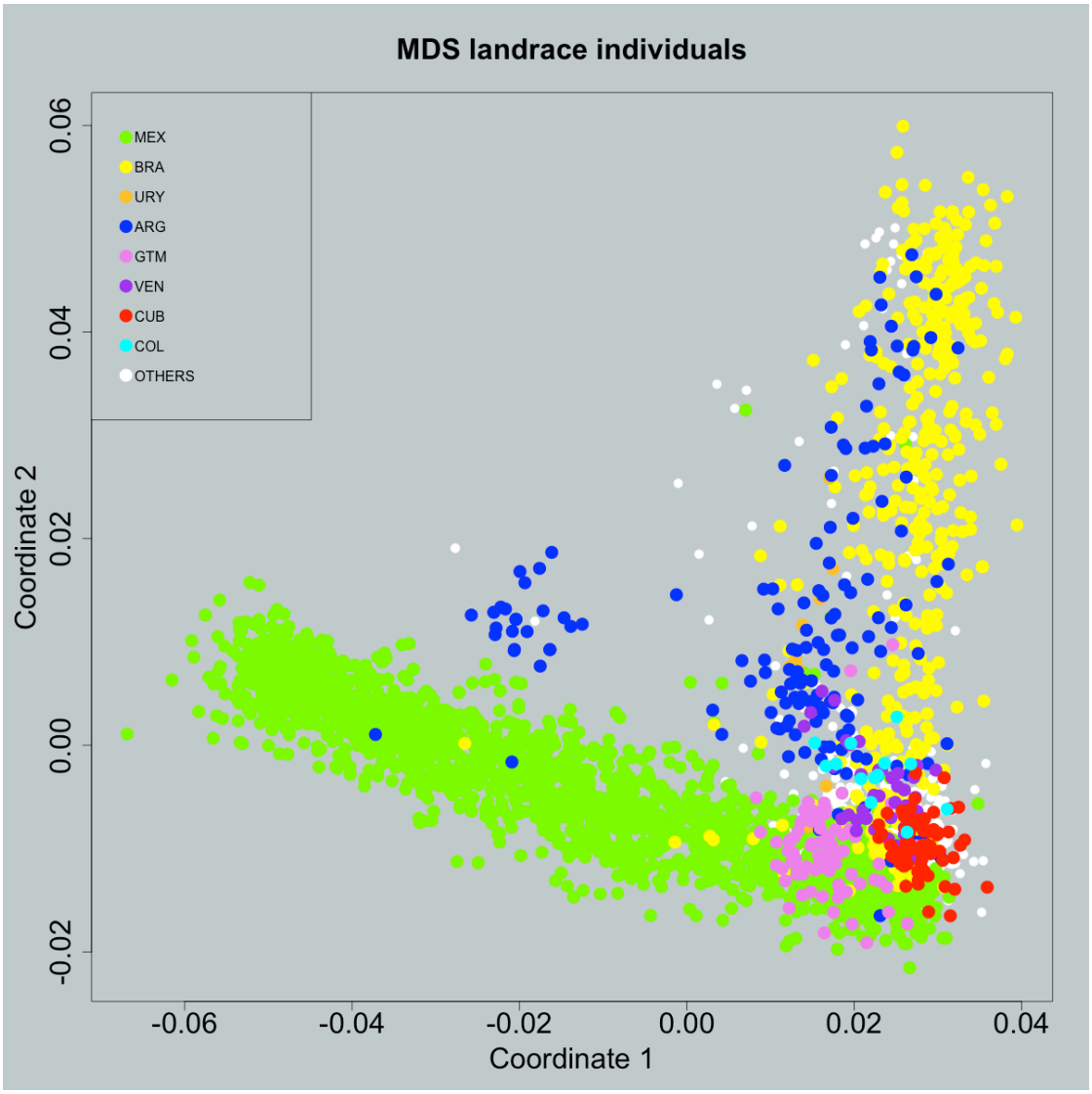


Figure 2.7: MDS of FOAM accessions

2.7 First 2 Principal Coordinates from Multidimensional scaling of the genetic distance among accessions

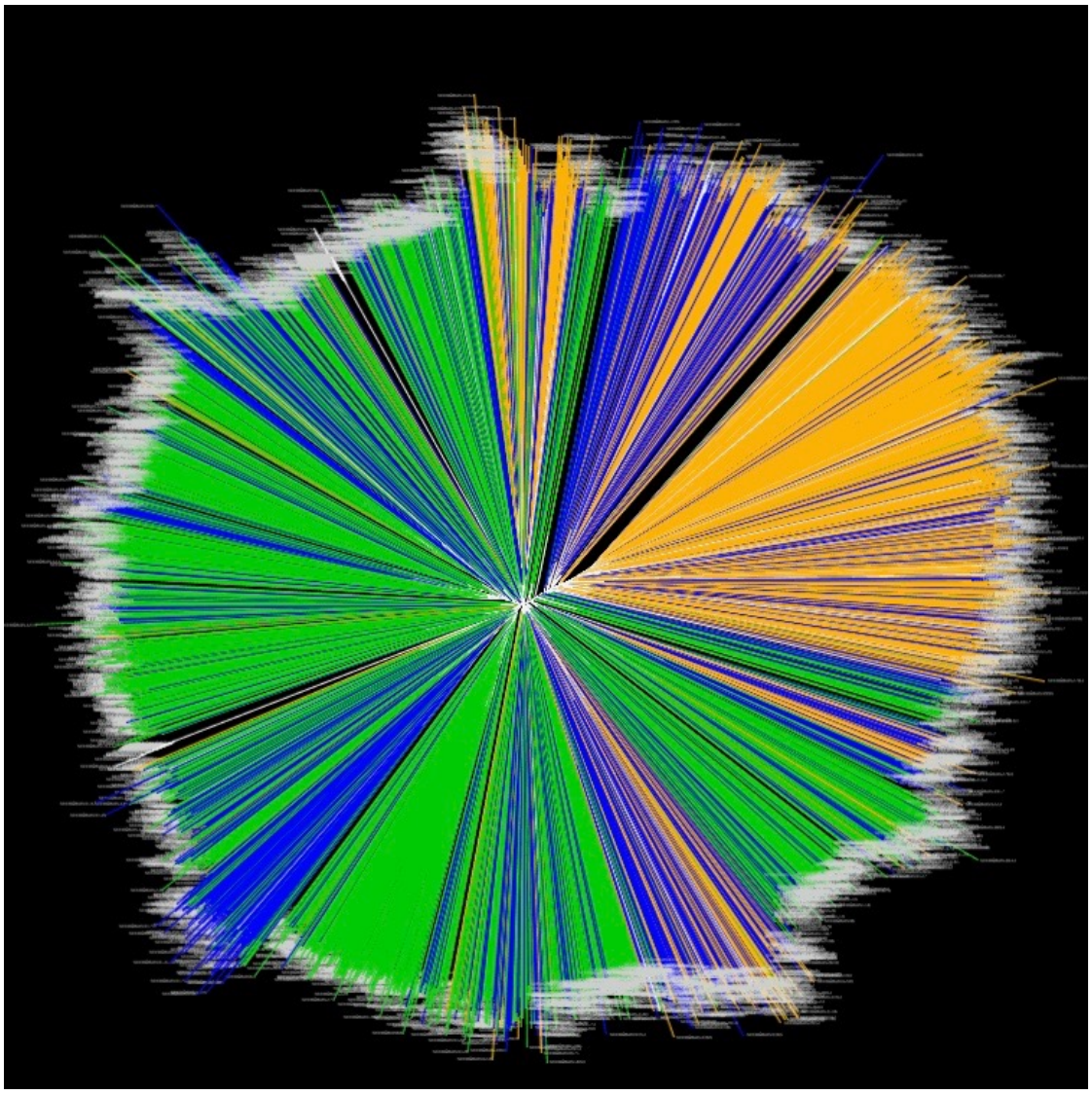


Figure 2.8: Neighbor-joining tree by adaptation class

2.8 Neighbor-joining tree of landrace individuals. Adaptation classes are colored green for low elevation, blue for mid elevation and orange for high elevation

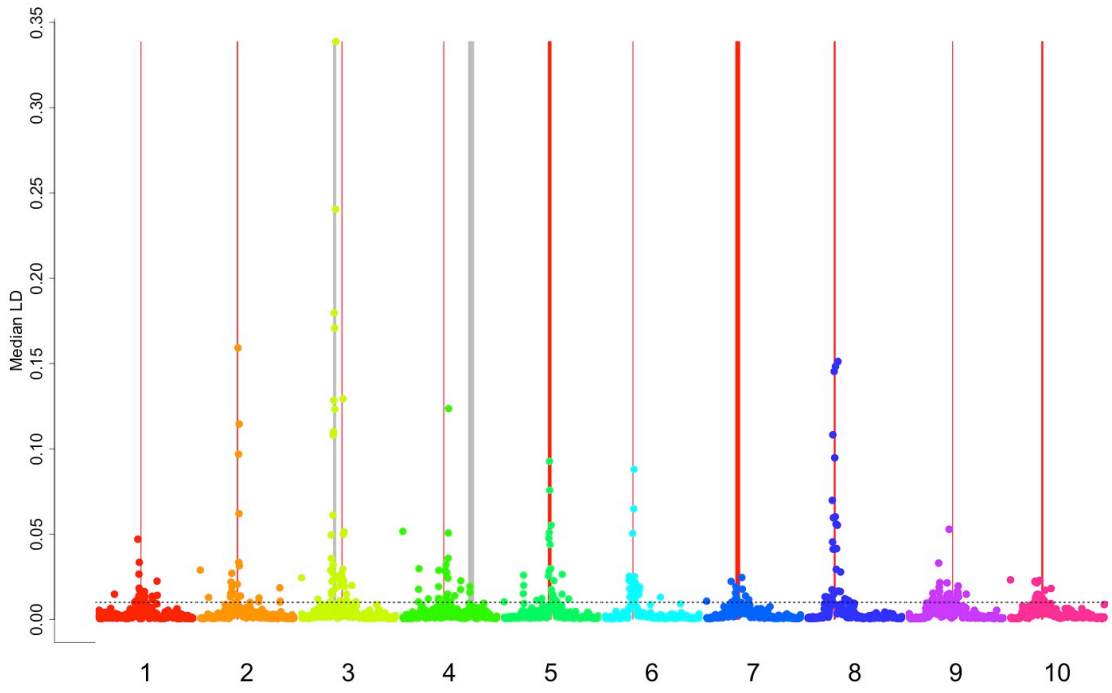


Figure 2.9: LD empirical threshold

2.9 Red shaded areas represent the centromeres, gray shaded areas represent inversions on chromosomes 3 and 4, and the dashed horizontal line represents the empirical LD threshold

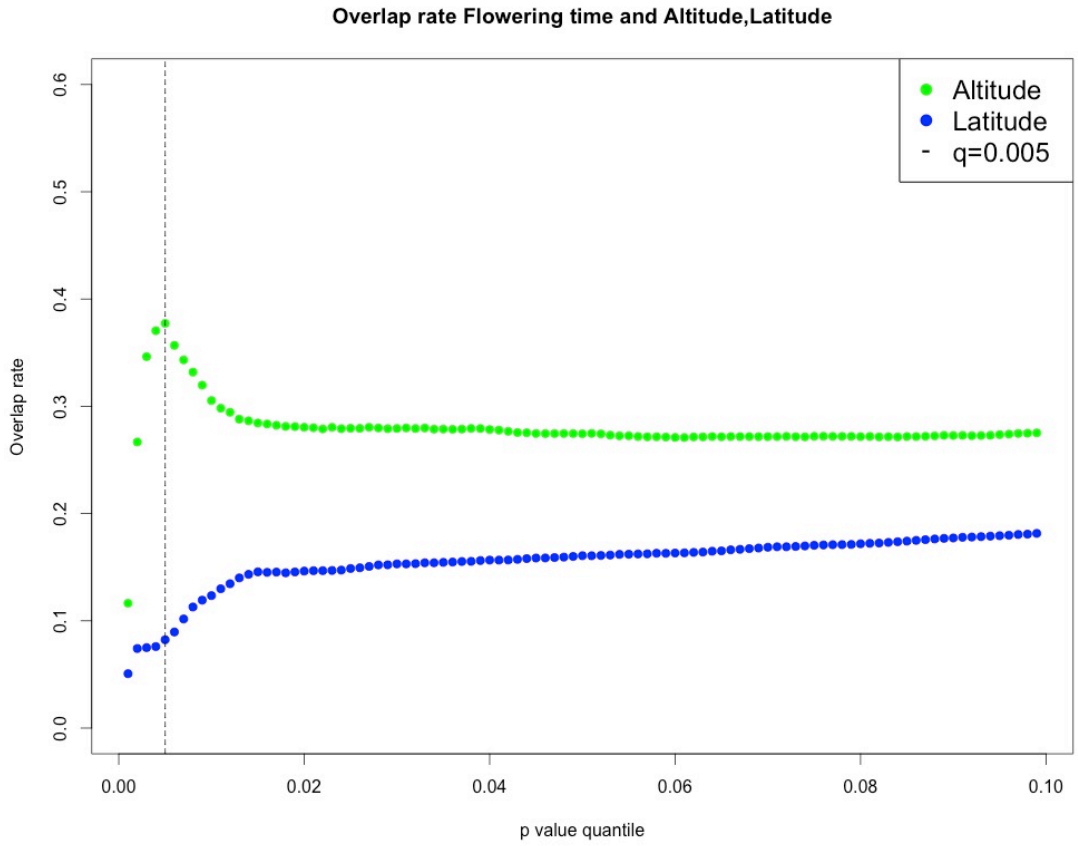


Figure 2.10: Overlap rate altitude and latitude with flowering time

2.10 Overlap rate between the top associating SNPs with flowering time and altitude, latitude at various p-value thresholds

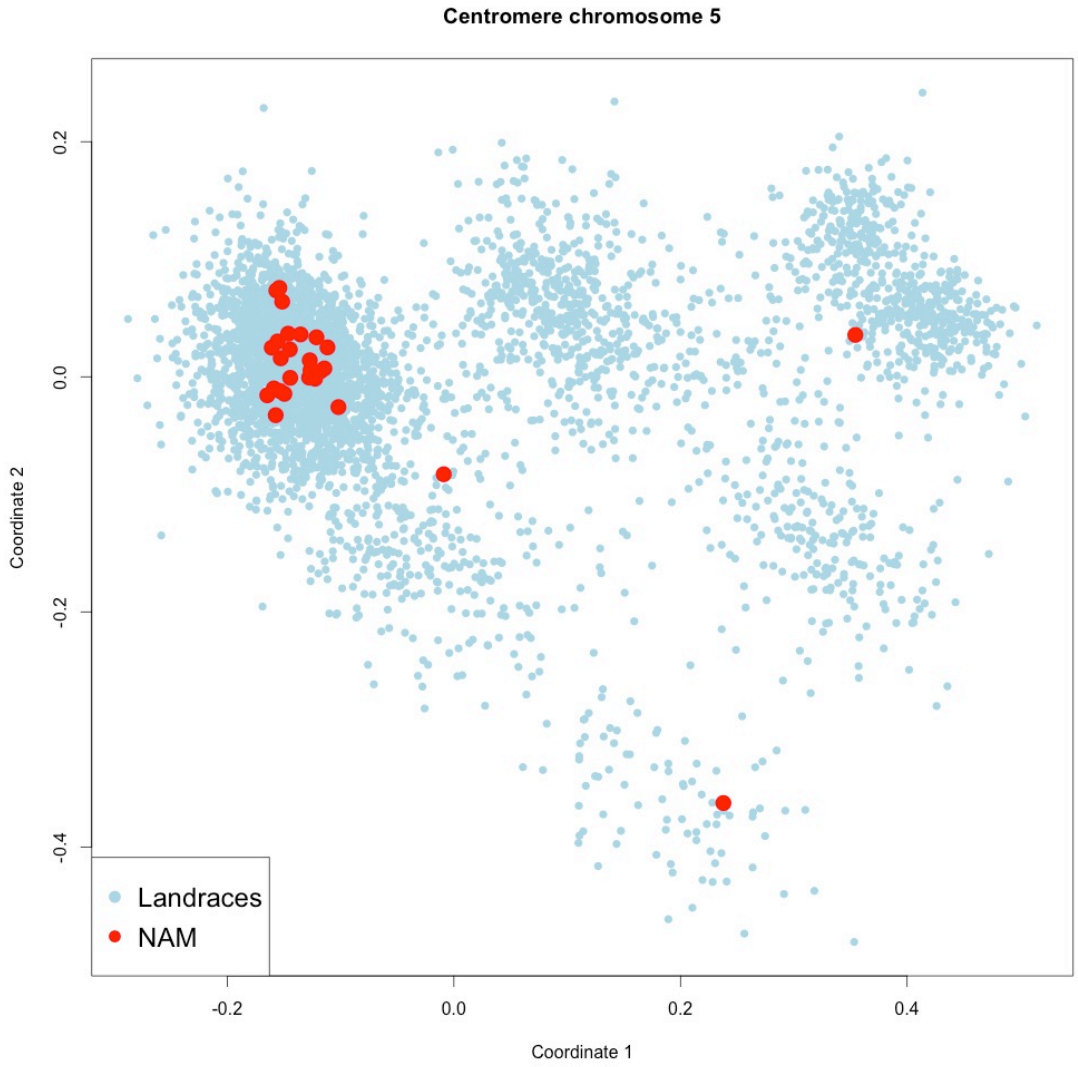


Figure 2.11: MDS of centromere of chromosome 5

2.11 MDS of centromere of chromosome 5 for the FOAM landrace accessions and the NAM founders: Topright: Il14H. Bottom: P39. Middle: CML333

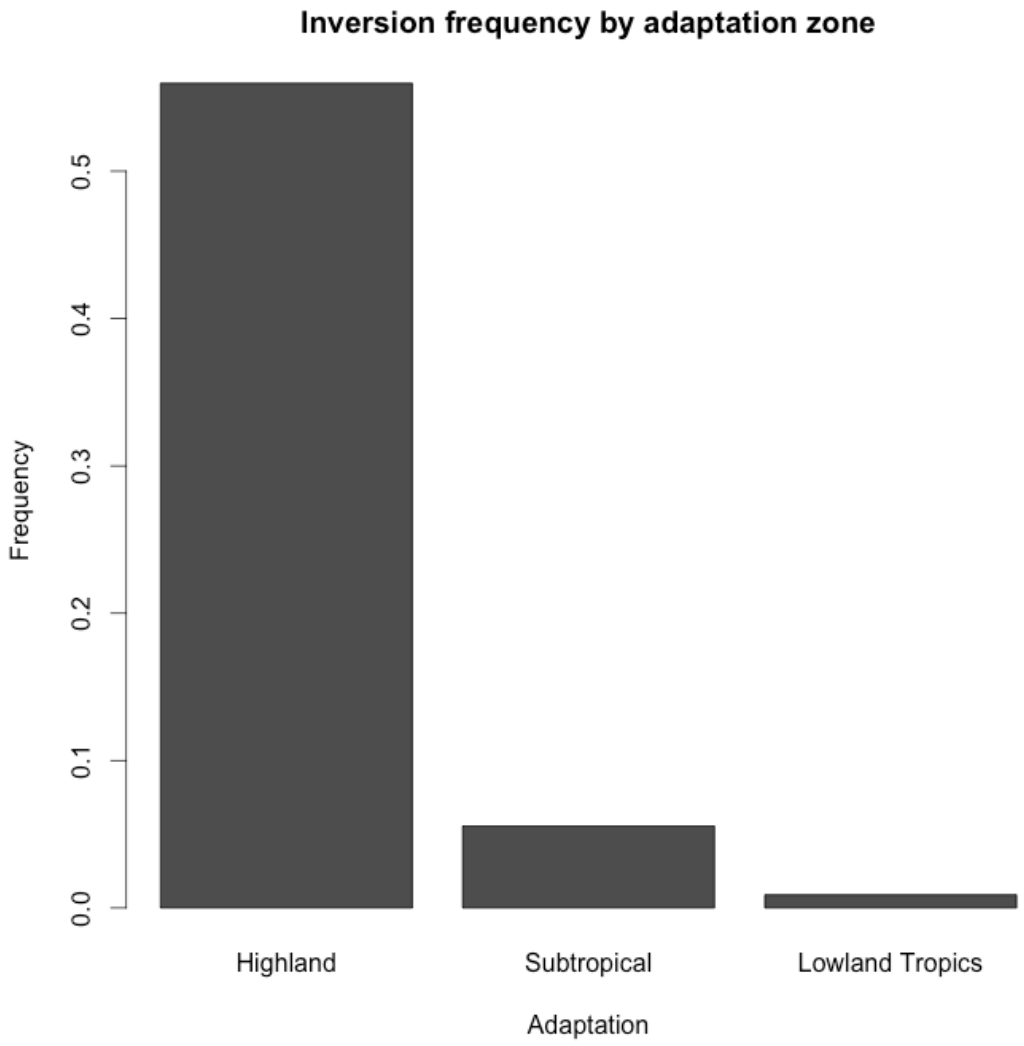
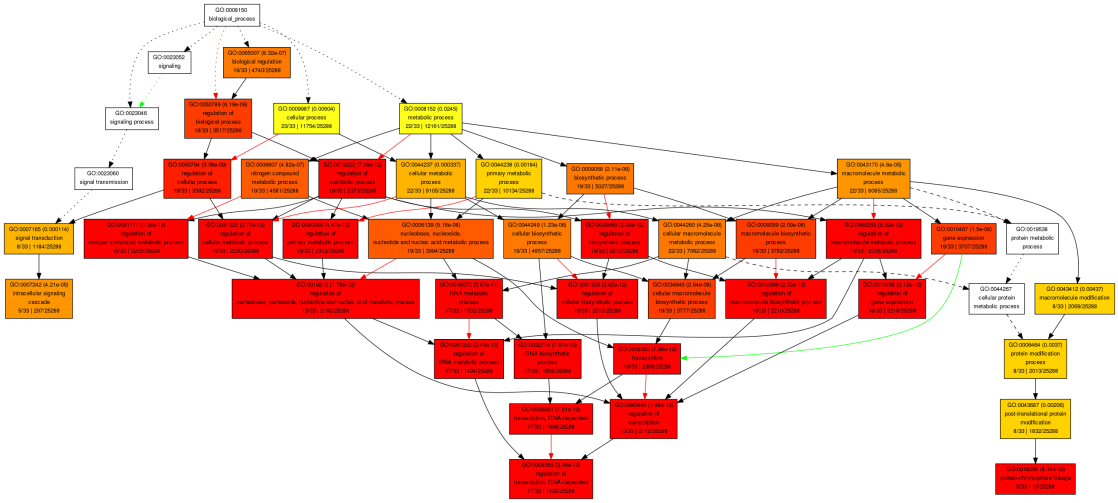
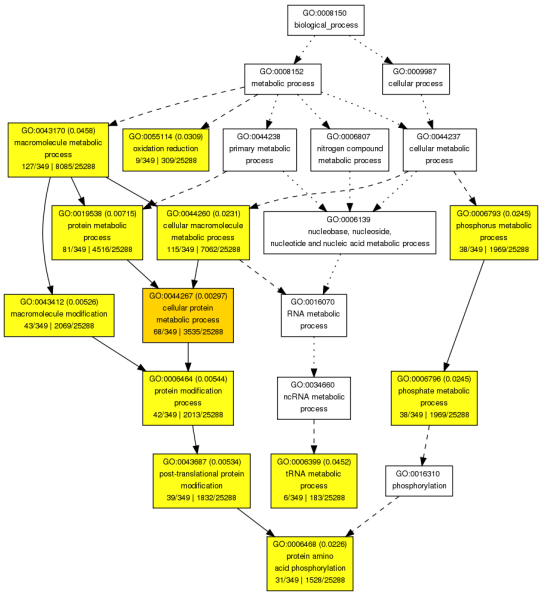


Figure 2.12: INV4 frequency by adaptation class

2.12 Frequency of INV4m according to accessions adaptation class



(a) Ontology for flowering time candidate genes



(b) Ontology for flowering time associating genes

Figure 2.13: Gene ontology enrichment of associating genes

2.13 Ontology of genes for a) previously reported flowering time genes and b) all the associating genes for this study, showing a significant enrichment of regulatory ontologies on the previously characterized genes, while our study has modest enrichment in general metabolism. Those genes probably have an indirect downstream effect on flowering time.

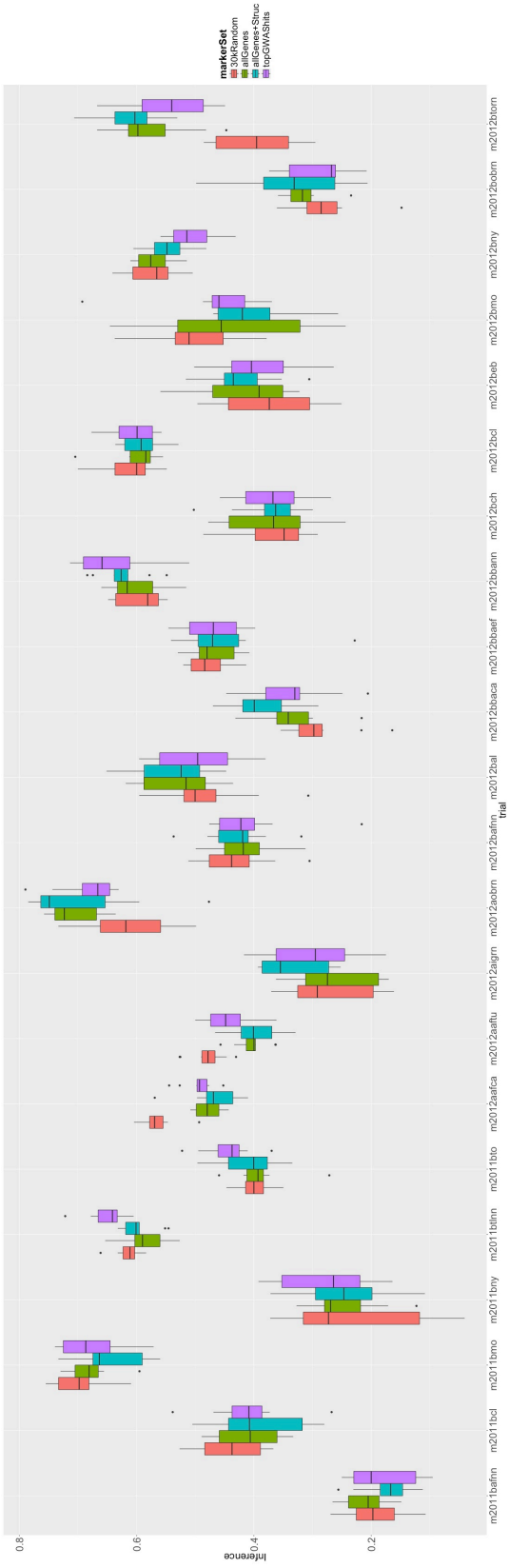


Figure 2.14: Genomic prediction accuracy by trial

Year/Cycle	Trial	Location	Adaptation	Accessions
2011B	m2011BAFnn	Agua Fria, Puebla	LOWLAND	1707
2011B	m2011BCL	Celaya, Guanajuato	SUBTROPICAL	783
2011B	m2011BMO	Tarimbaro, Michoacan	SUBTROPICAL	481
2011B	m2011BNY	San Pedro Lagunillas, Nayarit	SUBTROPICAL	551
2011B	m2011BTLnn	Tlaltizapan, Morelos	SUBTROPICAL	1140
2011B	m2011BTO	Torreon, Coahuila	SUBTROPICAL	1403
2012A	m2012AAFca	Agua Fria, Puebla	LOWLAND	1921
2012A	m2012AAFtu	Agua Fria, Puebla	LOWLAND	1923
2012A	m2012AIGrn	Iguala, Guerrero	LOWLAND	749
2012A	m2012AOBrn	Obregon, Sonora	LOWLAND	452
2012B	m2012BAFnn	Agua Fria, Puebla	LOWLAND	717
2012B	m2012BAL	Amoloya de Juarez, Mexico	HIGHLAND	428
2012B	m2012BBAca	El Batan, Mexico	HIGHLAND	817
2012B	m2012BBAef	El Batan, Mexico	HIGHLAND	759
2012B	m2012BBAnn	El Batan, Mexico	HIGHLAND	817
2012B	m2012BCH	Guadalupe-Victoria, Chiapas	LOWLAND	671
2012B	m2012BCL	Celaya, Guanajuato	SUBTROPICAL	805
2012B	m2012BEB	m2012BEB	SUBTROPICAL	658
2012B	m2012BMO	Numaran, Michoacan	SUBTROPICAL	282
2012B	m2012BNY	San Pedro Lagunillas, Nayarit	SUBTROPICAL	805
2012B	m2012BOBrn	Obregon, Sonora	LOWLAND	523
2012B	m2012BTOrn	Torreon, Coahuila	SUBTROPICAL	338

Table 2.1: Trial years, locations, and number of accessions

CHAPTER 3

GENOME-ENVIRONMENT ASSOCIATION ALLOWS IDENTIFYING USEFUL ADAPTIVE ALLELES FROM MAIZE LANDRACES

3.1 Abstract

Domesticated species represent some of the most outstanding cases of human directed evolution. Through recurrent selection, several plants and animals have been adapted to highly diverse environments, yet generally the genes relevant for each environmental condition remain uncharacterized. We studied the genetic architecture of adaptation to 2 soil and 10 climatic variables in a comprehensive panel of 3,312 maize landrace accessions from Latin America. We observed that overall 90% of the genome exhibited very low differentiation (F_{ST}=0.058) across altitudinal clades, with 26.9% of the altitudinal differentiation being the productive introgression from a distinct wild high elevation subspecies. In general across traits significant SNPs had high minor allele frequency, were shared across related traits, and on average 43% of the significant SNPs were contained in low recombination regions. Several genes were significant across climate traits, including an aldehyde dehydrogenase for soil pH, and in particular for many of the water use traits we observed significant association with the gene *hsftf9*, a heat shock transcription factor. We showed that variation within *hsftf9* displays a significant effect in drought trials in unrelated material, contributing to yield and flowering time under stressed conditions, highlighting the potential to use this approach for mining of useful alleles for adaptive traits from landraces. The genetic architecture of landrace local adap-

[0]For this chapter, my contribution encompassed the population differentiation analysis, the genome wide association models, and gene level analyses

tation can provide knowledge critical to adapt the worlds high production crop to to global climate change.

3.2 Results

Maize was domesticated in the tropical lowlands of the Balsas river in Mexico nearly 10,000 years before present[1], and since it has been adapted to highly divergent environmental conditions. The broad adaptation seen in maize today is likely the product of its ancestors being adapted to extremely diverse environments within central Mexico. Post domestication maize has moved across the globe to diverse environments becoming the worlds highest production food and feed crop. This broad adaptation was achieved thanks to the wide genetic variation and population history of the crop, incorporating useful alleles from its wild ancestor and related species[2,45,46], and allowing gene flow across populations through local pollen flow and broader seed exchange, creating a panmictic population that acts like a powerful genetic "mixing pot". Although some studies have analyzed specific components of adaptive traits in maize lines, the genetic basis of adaptation[36] to environmental conditions in maize traditional varieties, or landraces remains largely unknown. We studied the genetic architecture of various components relevant to abiotic stress by examining the association between genome wide variation with environmental and ecogeographic covariates. Our sample included 3,312 landrace individuals from CIMMYT global collections of landraces. These individuals were each sampled from one population, and together reflect the landrace diversity from 36 countries in Latin America. We used the available high density genotyping data[44] and collection site information on these landraces to study climate and soil traits.

We first estimated the differentiation among landraces accessions to see the extent of allele sharing across environments. Previous studies in Latin American maize landraces[7] have shown that altitude is an important barrier for the exchange of alleles and three adaptation clines have been identified. We estimated the fixation index (F_{ST}) across the three previously defined altitude classes. We observed a significant overlap between SNPs with high F_{ST}, regions with extended LD, and variation corresponding to recent introgressions from *Zea mays* ssp. *mexicana* into highland maize[2] including the centromeres 1, 5, 6, and 10, INV4, two regions on chromosome 9, and one on chromosome 2. These regions agree with the regions associated with altitude adaptation in the sample. Outside those regions, we observed that 90% of the genome had fixation index values under 0.058, indicating that only a small portion of the genome displays strong differentiation, mostly alleles relevant to altitude adaptation, while the alleles on the rest of the genome segregate across elevation gradients. This in turn has as consequence a uniquely high power and resolution for resolving causative SNPs and studying local adaptation.

Recent studies have shown the potential to use genome-environment association to study local adaptation[8]. Therefore, we performed genome wide association (Methods) between our high density markers with the following environmental variables adjusted for a 6 month growing season: for climate, we looked at precipitation[9,10]; cloud cover[10]; monthly mean daily average, minimum and maximum temperature[9]; diurnal temperature range[10]; frost frequency[10]; potential evapotranspiration[11]; wet day frequency[10]; and vapor pressure[10]. In addition for soil we studied adaptation to differences in pH[12] and waterlogging[13]. We observed that on average across traits 43% of the significant SNPs were contained within *Zea mays* ssp. *mexicana* introgressions (3.1) with the highest and

lowest (78%, 7.5%) corresponding to mean daily temperature and precipitation respectively. This suggests that for environmental traits associated with altitude, landraces obtained a significant fraction of the useful alleles from its sister species.

When looking jointly at the minor allele frequency distribution of the associating SNPs, we observed that unlike the null distribution of genetic variation, significantly associated SNPs are high in frequency. This is in contrast to flowering time, an adaptive trait analyzed on the same population[44] which displayed enrichment at low frequency for significant SNPs. This suggests that across their wide geographic range, maize landraces have been matched with their environmental conditions through selection at the same loci, with the segregation of functional adaptive alleles being maintained across populations.

We looked at the significant genes outside the high-LD regions, and we observed that several genes showed association with more than one environmental trait. In particular, 13 genes showed significant association for 7 climatic traits, and annotation was available for 11 of their *Arabidopsis* homologs (3.2). In addition, we looked at overlap between associating SNPs and a list of 152 dehydration responsive element binding (DREB) transcription factors proteins[14], a family with known relevance in plants response to a variety of abiotic stresses[15]. Wet day frequency and precipitation had the highest overlap with DREB genes, with 14 and 11 significant DREBs each, followed by potential evapotranspiration and frost frequency with 6 and 4. Although gene ontology enrichment within each climate trait varied, the ontology of all the traits was significantly enriched for biological regulation, and regulation of transcription (FDR p value 7.20E-10, Figure 3.3).

We observed a low correlation between cloud cover and the other environmental traits, with the highest corresponding to potential evapotranspiration (r=0.42, Figure 3.2). The most significant gene associated with cloud cover trait was GRMZM2G036980 (Figures 3.10,3.9). This gene contains a VQ motif, and its best hit in rice is a putatively expressed VQ domain containing protein. There are 34 proteins containing the VQ motif in *Arabidopsis*, some of which are associated with growth and disease response phenotypes, and potentially acting as cofactors of WRKY transcription factors[16]. For frost frequency, we observed a modest correlation with minimum, maximum and average temperature range (r=0.5, Figure 3.2) and we observed prevalence of frost frequency to be highest in the Andean region of South America. The most significant gene associated with frost frequency was AC185415.3_FG005 (Figures 3.5, 3.6), which contains an RNA recognition motif (RRM) and two zinc finger domains. Its best hit in rice is annotated as zinc finger RNA-binding domain-containing protein 2, and in *Arabidopsis* its best hit is the TBP-associated factor 15 (TAF15). There are 18 TAF in *Arabidopsis*[17] and at least one is adaptive under salt stress[18], and more broadly RNA-binding proteins have been associated to be responsive to a variety of stresses in *Arabidopsis* including freezing tolerance[19]

We observed high correlation between minimum, maximum, and average daily temperature (Figure 3.2). The most significant gene associated with minimum and average daily temperature was GRMZM2G001265 (3.21, Figures 3.22, 3.23, 3.24), with its gene product annotated in maize as palmitoyltransferase ZDHHC20, and its best hit in rice being being a DHHC zinc finger domain containing protein. The best hit corresponding to this gene in *Arabidopsis* is PAT14, which encodes an S-acyltransferase involved in the regulation of leaf senescence[20]. For maximum average daily temperature, the most significant

hit was GRMZM2G375856, which contains the WD40 domain and is expressed only in anthers in B73. The best hit associated to this gene in *Arabidopsis* is AT3G15470, which in turn is annotated as a member of the Transducin/WD40 repeat-like superfamily. This domain is present in several proteins across eukaryotes, with no known catalytic activity but prevalent interaction with other proteins[21], and at least one member of this family in *Arabidopsis* has been associated with climate variation in that organism[22].

We were particularly interested in traits related to water use, as it is an important factor for crop productivity, and for global food security it will be critical to deploy alleles matching the changes in precipitation and weather[23,24]. We observed a region on chromosome 9 that consistently showed association for several environmental traits related to water use, including diurnal temperature range (Figures 3.11, 3.12), potential evapotranspiration (Figures 3.13, 3.14), precipitation (Figures 3.15, 3.16), vapor pressure (Figures 3.17, 3.18) and wet day frequency (Figures 3.19, 3.20), even with the traits having modest correlation (r=0.6, Figure 3.2). This region contained nine genes, with significant markers around two clusters containing three and two genes respectively. The genes at the first cluster have the highest significance across all traits, and include first GRMZM2G392737, a putative MAP kinase with a rice best hit corresponding to a STE kinase and a best hit in *Arabidopsis* corresponding to MAP kinase kinase7. The second gene, GRMZM2G093254, contains the conserved domain Sec-independent protein translocase protein (TatC). Its best hit in *Arabidopsis*, AT2G01110, also known as ALBINO AND PALE GREEN 2, is annotated as a core subunit of the chloroplast twin-arginine translocation (Tat) translocase, which is present in the thylakoid membrane and is directly involved in the transport of folded proteins[25]. The third gene in the cluster is

GRMZM2G528283, which has no best hits in sorghum, rice or *Arabidopsis*; has no known domains; and contains one transposable element insertion, probably corresponding to a non-functional gene. The second cluster with significant association with climate traits includes first the gene GRMZM2G328268. This gene contains the Structural Maintenance of Chromosomes (SMC) domain, corresponding to a Chromosome segregation ATPase and has no annotated hits in *Arabidopsis*, rice or sorghum. The second gene in the region, GRMZM2G026742 is annotated as *hsftf9*, a heat shock factor protein homologous to *Arabidopsis* HSF1A. Because these genes were significant for diurnal temperature range, potential evapotranspiration, vapor pressure, precipitation and wet day frequency, we hypothesized a role for plants under water stress. In order to validate the potential effect of *hsftf9*, we looked at its effect on yield, days to anthesis, and the anthesis silking interval in the unrelated Drought Tolerant Maize for Africa panel [26,27]. We used phenotypic data evaluated under regular, drought, high heat, and combined heat and drought management regimes. We observed one SNP within the gene *hsftf9* to be significantly associated with yield (pval= 1.554e-05) and flowering time (pval <2e-16). Compared to the irrigated control, this SNP displayed significant genotype by environment interaction only for yield under drought and drought with heat stress(pval= 0.001).

In contrast to climatic traits, adaptation to pH and waterlogging displayed association with very few genes (-log10 p >15). In total we observed 3 regions of the maize genome associated to changes in pH. The genes include on chromosome 4 (Figures 3.26,3.25) the gene GRMZM2G123667, in maize also annotated as *nactf125*, which is part of the extensive family of plant NAC transcription factors. NAC transcription factors have been implicated in a very wide range of biological functions, including abiotic stress response[28]. In close

proximity to *nactf125*, we observed significant association with the gene GRMZM2G060800, which is homologous to the *Arabidopsis* aldehyde dehydrogenase ALD3H1. Aldehyde dehydrogenase enzymes are widespread and highly conserved proteins, some displaying an important role in osmotic stress response, as well as adaptation to dehydration, salt, heavy metals, methyl viologen, and H2O2 [29,30]. Adjacent to that gene (800bp) is GRMZM2G060898, which contains the Cytochrome b domain, and its rice and *Arabidopsis* best hit is annotated as encoding the cytochrome b(6) subunit of the cytochrome b6f complex. Finally, the significant gene GRMZM2G702152 is a homolog to *Arabidopsis* early in short days4 (*esd4*), a protease involved in the covalent attachment of small ubiquitin-like modifier (SUMO)[31]. SUMO proteins are found in several species, and have been implicated in stress response[32,33]. The 4 genes associated with waterlogging were GRMZM2G065420, AC205405.3_FG001, GRMZM2G424147, and GRMZM5G827026. Only the last two share homology with *Arabidopsis*, with GRMZM2G424147 being homologous to AT1G02080, a transcriptional regulator which in *Arabidopsis* displays modified expression upon salt stress downstream of AtNAC2, a NAC-type transcription factor gene[34]. GRMZM5G827026 is homologous to a Polyketide cyclase/dehydrase and lipid transport superfamily protein, and recent reports suggest that members of this gene family are implicated in wounding response[35].

Crop landraces have been selected through recurrent selection by farmers for local adaptation over hundreds or even thousands of generations. More recently, important efforts have allowed the collection of extensive collections for multiple species, with detailed information on sampling location. The geographic associations here reported exemplify the potential of combining high-density genotyping along with climatic data from landraces to mine useful al-

leles relevant for adaptive traits. The high resolution and power achieved in this mapping study is partly due to the high frequency of the adaptive loci, along with the very low linkage disequilibrium in maize. In addition, we observed a similar phenomenon to humans. In *homo sapiens* introgressions from now extinct homonins[47] likely played an important role in local adaptation with alleles relevant to traits as diverse as skin pigmentation, metabolism, immune response, and adaptation to high elevation. Similarly, in maize we observed significant contribution of at least one structural introgression, INV4, which has been previously associated to adaptation to high elevation, and which together with other regions that display high linkage disequilibrium are associated to adaptation to several climatic traits. With increasing climate variability, having a catalog of useful alleles for specific environmental conditions will become critical to ensure the resilience of crops, and could help guide the search for adaptive alleles in other species.

3.3 Methods

Germplasm and environmental covariates

We used individuals from 3,312 landrace accessions from CIMMYT global collection. These were collected from 1947 to 2007, with a mean collection year of 1965 and 90% of the samples having a collection date prior to 1980. Therefore, these landraces reflect the diversity prior to the large scale use of improved hybrid varieties in Latin America starting in the 1980s and 1990s. Latitude, longitude and elevation were recorded during genetic field sampling. Soil properties and classification data (e.g., pH, soil taxonomy) were from the Harmonized

World Soil program, a comprehensive database that combines all available regional soil databases with global soil maps (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012). Climate and weather-related data were obtained from various sources. Data for daily, monthly, and yearly averages of precipitation and temperature were from the WorldClim database (http://www.worldclim.org), a dataset generated from a large collection of global weather stations during the period 1950-2000 (Hijmans et al., 2005). Diurnal temperature ranges, frost frequency, wet day frequency, vapor pressure, and cloud cover data were obtained from the global CRU TS3.10 dataset, a high-resolution global grid of monthly meteorological observations (Harris et al., 2014). Potential evapotranspiration, mean annual precipitation and mean annual evapotranspiration were obtained from global models of annual evapotranspiration and precipitation (Zomer et al., 2008; Trabucco et al., 2008) and WorldClim precipitation and temperature data (Hijmans et al., 2005). Poorly-drained soils, indicated by the percentage of an area with waterlogged soils, were acquired from maps generated by the Soil Functional Capacity Classification System (Sanchez et al. 2003, HarvestChoice 2010). All spatial data were processed using ArcGIS 10.2.2 software (ESRI, Redlands, CA).

Genetic markers

For the landrace accessions, we obtained the BEAGLE4[36] and non-imputed GBS markers[44] reported in the flowering time chapter. For the DTMA, we obtained GBS markers imputed using FILLIN[37]. Both marker datasets are expressed in the coordinate system corresponding to the second version of the maize B73 reference genome. For association with climate traits, BEAGLE4 imputed markers were used, after filtering for minor allele frequency greater than 1%. A random

sample of 30,000 random SNPs was used to estimate 2 principal coordinates for accounting for population structure in the PC model.

Estimation of F_{ST}

Fixation index was estimated using the software vcftools[38] using non-imputed GBS markers with per site missing data less than 0.4. The 3 population groups were defined according to their adaptation class as lowland tropical (low elevation, <1200 meters above sea level and <30 N or 40 S), subtropical (mid elevation, between 1200 and 1900 m.a.s.l. and <30 N or 40 S), and tropical highland (above 1900 m.a.s.l. and <30 N or 40 S). The top percentile, corresponding to an F_{ST} value of 0.21, was used as threshold for significant differentiation. Regions corresponding to introgressions from *Zea mays* ssp. *mexicana* were obtained from Hufford et al, 2013 Table S4 column regionOfIntrogression.

Genome Wide Association

As populations adapt to different environments, the alleles underlying local adaptation increase in frequency to match their genetic effect with the corresponding environment, maximizing population fitness. However, many other alleles unrelated to local adaptation can be expected to show a similar differentiation across the various populations and would be related only to population. This confounding relationship has as consequence that models accounting for population structure decrease the false positive rate of associations, but in doing so also decrease significantly the ability to detect true positives. We performed Genome wide association for each trait with 3 independent generalized linear

models using the software TASSEL[39], and by comparing the results from the different models we aim to differentiate between alleles involved in local adaptation and those involved in population structure. The models used were 1) a model with no covariates, which we referred to as naive model, 2) a model including as fixed effects the first 2 principal coordinates from Multidimensional scaling to account for population structure, and 3) a model including as fixed effects the altitude, latitude, and longitude sampling coordinates.

Each model has very specific strengths and drawbacks related to true vs false positives. The significant results from model 1) will include both population structure and local adaptation related alleles. In the case of model 2) we used only 2 MDS weights because in the MDS space represented by such weights, landraces display a pattern that generally matches their genetic origin, including the more recent introductions from Northern Mexican and Southern North American landraces to some South American locations. The signficant results from this model would decrease false positives, however alleles involved in local adaptation correlated with those 2 MDS weights would be expected to show limited or null significance. Finally, we used model 3) to account directly for geography-driven population structure. An additional concern then becomes the threshold for considering SNPs from each model as significant. The p-value distributions from all models display significant deviation from the expectation, however the use of mixed linear models decreases genomewide significance all together. In order to establish a threshold for significance, we decided to incorporate evidence from a phenotype correlated with local adaptation. We used the trait of flowering time, for which phenotypic evaluation was collected, and for which we expected significant overlap between alleles involved in flowering time and alleles relevant for climatic adaptation. In order to consider those

results together, we estimated a measure we call ”overlap rate” between the top SNPs from flowering time and the top SNPs for each trait for each of the 3 GWAS models. Overlap rate was defined as the sum of the overlapping SNPs between the top flowering time SNPs from this panel and the significant SNPs for each model across significance quantiles from 0.001 to 0.05 divided by the union of the sets. The p-value threshold was chosen to maximize overlap rate curve. Once p-value thresholds were obtained for each model for each climate trait, overlapping SNPs bewteen the two models with the highest agreement were chosen. We chose the two models with highest agreement for all traits except for wet day frequency, for which only the naive model displayed overlap with flowering time, with the models accounting for population structure showing no overlap.

Unlike climatic traits, pH and waterlog percent do not show overlap with flowering time associated genes, therefore overlap rate with flowering time was not used for these traits. These traits also displayed little deviation from expected p-value distribution due to population structure, therefore for both soil traits only the SNPs significant for both models with covariates were considered, and based on the Manhattan plots the p value threshold chosen was -log10 p value greater than 15.

For the analyses on the DTMA panel, previously published BLUPs for days to anthesis, yield, and the anthesis silking interval were obtained[26]. We used the Genotyping by Sequencing SNP markers available for the DTMA lines. The 21 markers around *hsfp9* (within 100kb) were each fitted independently as explanatory variable in linear regressions for each trait. The model included a main effect for irrigation, marker, and their corresponding interaction term. For

gene ontology enrichment, we used the gene ontology enrichment tool agriGO[40]

Significant genes

In order to obtain significant genes, we first created a set of regions with extended LD. This set was the combination of the previously reported high-LD regions on the same panel, all the centromeres, and the coordinates corresponding to INV4. In total, this set encompasses 128 Mb, roughly 5% of the of the maize genome. In order to find associating genes, only significant SNPs outside of the high-LD set are matched to the nearest annotated gene. We used the GenomicRanges package in R [41] to match the associating SNPs, the high-LD region, as well as the annotated genome from the general feature format (gff) file containing the filtered gene set on reference genome version 2, which matches the coordinate system of the genetic markers. The gff file was filtered to consider only the 39,249 protein coding genes. Gene information was obtained from MaizeGDB[42,43].

3.4 References

1. Piperno, D. R., Ranere, A. J., Holst, I., Iriarte, J. & Dickau, R. Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas.

2. Hufford, M. B. et al. The genomic signature of crop-wild introgression in maize. PLoS Genet. 9, e1003477 (2013).

3. Takuno, S. et al. Independent molecular basis of convergent highland adaptation in maize. bioRxiv (2015). doi:10.1101/013607

4. Bertin, P., Madur, D., Combes, V., Dumas, F. & Brunel, D. Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with a . PLoS One (2013).

5. Hayes, K. R. et al. Maize global transcriptomics reveals pervasive leaf diurnal rhythms but rhythms in developing ears are largely limited to the core oscillator. PLoS One 5, e12887 (2010).

6. Chia, J.-M. et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44, 803807 (2012).

7. Salhuana, W., Jones, Q. & Sevilla, R. The Latin American Maize Project: Model for rescue and use of irreplaceable germplasm. Diversity (1991).

8. Lasky, J. R. et al. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv 1, e1400218 (2015).

9. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 19651978 (2005).

10. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly climatic observations the CRU TS3.10 Dataset. Int. J. Climatol. 34, 623642 (2014).

11. Zomer, R. J., Trabucco, A., Bossio, D. A. & Verchot, L. V. Climate change mitigation: A spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agric. Ecosyst. Environ. 126, 6780 (2008).

12. Nachtergaele, F. & Batjes, N. Harmonized world soil database. (FAO Rome, Italy, 2012).

13. Sanchez, P. A., Palm, C. A. & Buol, S. W. Fertility capability soil classification: A tool to help assess soil quality in the tropics. Geoderma. Geoderma 114,

157185 (2003).

14. Liu, S. et al. Genome-wide analysis of ZmDREB genes and their association with natural variation in drought tolerance at seedling stage of Zea mays L. PLoS Genet. 9, e1003790 (2013).

15. Lata, C. & Prasad, M. Role of DREBs in regulation of abiotic stress responses in plants. J. Exp. Bot. 62, 47314748 (2011).

16. Cheng, Y. et al. Structural and functional analysis of VQ motif-containing proteins in Arabidopsis as interacting proteins of WRKY transcription factors. Plant Physiol. 159, 810825 (2012).

17. Lago, C., Clerici, E., Mizzi, L., Colombo, L. & Kater, M. M. TBP-associated factors in Arabidopsis. Gene 342, 231241 (2004).

18. Gao, X., Ren, F. & Lu, Y.-T. The Arabidopsis mutant stg1 identifies a function for TBP-associated factor 10 in plant osmotic stress adaptation. Plant Cell Physiol. 47, 12851294 (2006).

19. Lorkovi, Z. J. Role of plant RNA-binding proteins in development, stress response and genome organization. Trends Plant Sci. 14, 229236 (2009).

20. Lai, J. et al. Two homologous protein S-acyltransferases, PAT13 and PAT14, cooperatively regulate leaf senescence in Arabidopsis. J. Exp. Bot. 66, 63456353 (2015).

21. Stirnimann, C. U., Petsalaki, E., Russell, R. B. & Mller, C. W. WD40 proteins propel cellular networks. Trends Biochem. Sci. 35, 565574 (2010).

22. Lasky, J. R. et al. Characterizing genomic variation of Arabidopsis thaliana: the roles of geography and climate. Mol. Ecol. 21, 55125529 (2012).

23. Wheeler, T. & von Braun, J. Climate change impacts on global food security. Science 341, 508513 (2013).

24. Elliott, J. et al. Constraints and potentials of future irrigation water avail-

ability on agricultural production under climate change. Proc. Natl. Acad. Sci. U. S. A. 111, 32393244 (2014).

25. Lee, P. A., Tullman-Ercek, D. & Georgiou, G. The bacterial twin-arginine translocation pathway. Annu. Rev. Microbiol. 60, 373395 (2006).

26. Cairns, J. E. et al. Identification of Drought, Heat, and Combined Drought and Heat Tolerant Donors in Maize. Crop Sci. 53, 1335 (2013).

27. Wen, W. et al. Molecular Characterization of a Diverse Maize Inbred Line Collection and its Potential Utilization for Stress Tolerance Improvement. Crop Sci. 51, 2569 (2011).

28. Olsen, A. N., Ernst, H. A., Leggio, L. L. & Skriver, K. NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. 10, 7987 (2005).

29. Stiti, N., Missihoun, T. D., Kotchoni, S. O., Kirch, H.-H. & Bartels, D. Aldehyde Dehydrogenases in Arabidopsis thaliana: Biochemical Requirements, Metabolic Pathways, and Functional Analysis. Front. Plant Sci. 2, 65 (2011).

30. Kirch, H.-H., Bartels, D., Wei, Y., Schnable, P. S. & Wood, A. J. The ALDH gene superfamily of Arabidopsis. Trends Plant Sci. 9, 371377 (2004).

31. Murtas, G. et al. A nuclear protease required for flowering-time regulation in Arabidopsis reduces the abundance of SMALL UBIQUITIN-RELATED MODIFIER conjugates. Plant Cell 15, 23082319 (2003).

32. Miller, M. J., Barrett-Wilt, G. A., Hua, Z. & Vierstra, R. D. Proteomic analyses identify a diverse array of nuclear processes affected by small ubiquitin-like modifier conjugation in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 107, 1651216517 (2010).

33. Hickey, C. M., Wilson, N. R. & Hochstrasser, M. Function and regulation of SUMO proteases. Nat. Rev. Mol. Cell Biol. 13, 755766 (2012).

34. He, X.-J. et al. AtNAC2, a transcription factor downstream of ethylene and

auxin signaling pathways, is involved in salt stress response and lateral root development. Plant J. 44, 903916 (2005).

35. Satheesh, V. et al. A Polyketide cyclase/dehydrase and lipid transport superfamily gene of Arabidopsis and its orthologue of chickpea exhibit rapid response to wounding. Ind. Jrnl. Gen. Plnt. Bree. 74, 463 (2014).

36. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459471 (2013).

37. Swarts, K., Li, H., Romero Navarro, J. A. & An, D. Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant 7, (2014).

38. Danecek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 21562158 (2011).

39. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 26332635 (2007).

40. Du, Z., Zhou, X., Ling, Y., Zhang, Z. & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 38, W6470 (2010).

41. Lawrence, M. et al. Software for computing and annotating genomic ranges. PLoS Comput. Biol. 9, e1003118 (2013).

42. Andorf, C. M. et al. MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res. 44, D1195201 (2016).

43. Andorf, C. M. et al. The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. Bioinformatics 26, 434436 (2010).

44. Romero Navarro J. Alberto, et al. Identifying the diamond in the rough: a study of allelic variation for flowering time in maize landraces (under review)

45. Doebley, J., Goodman, M.M. & Stuber, C.W. Patterns of isozyme varia-

tion between maize and Mexican annual teosinte Econ Bot (1987) 41: 234 46. Warburton, M. L. et al. Genetic Diversity in CIMMYT Nontemperate Maize Germplasm: Landraces, Open Pollinated Varieties, and Inbred Lines. Crop Sci. 48, 617 (2008).

47. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Snchez, E. Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. 16, 359371 (2015).

3.5 Figures

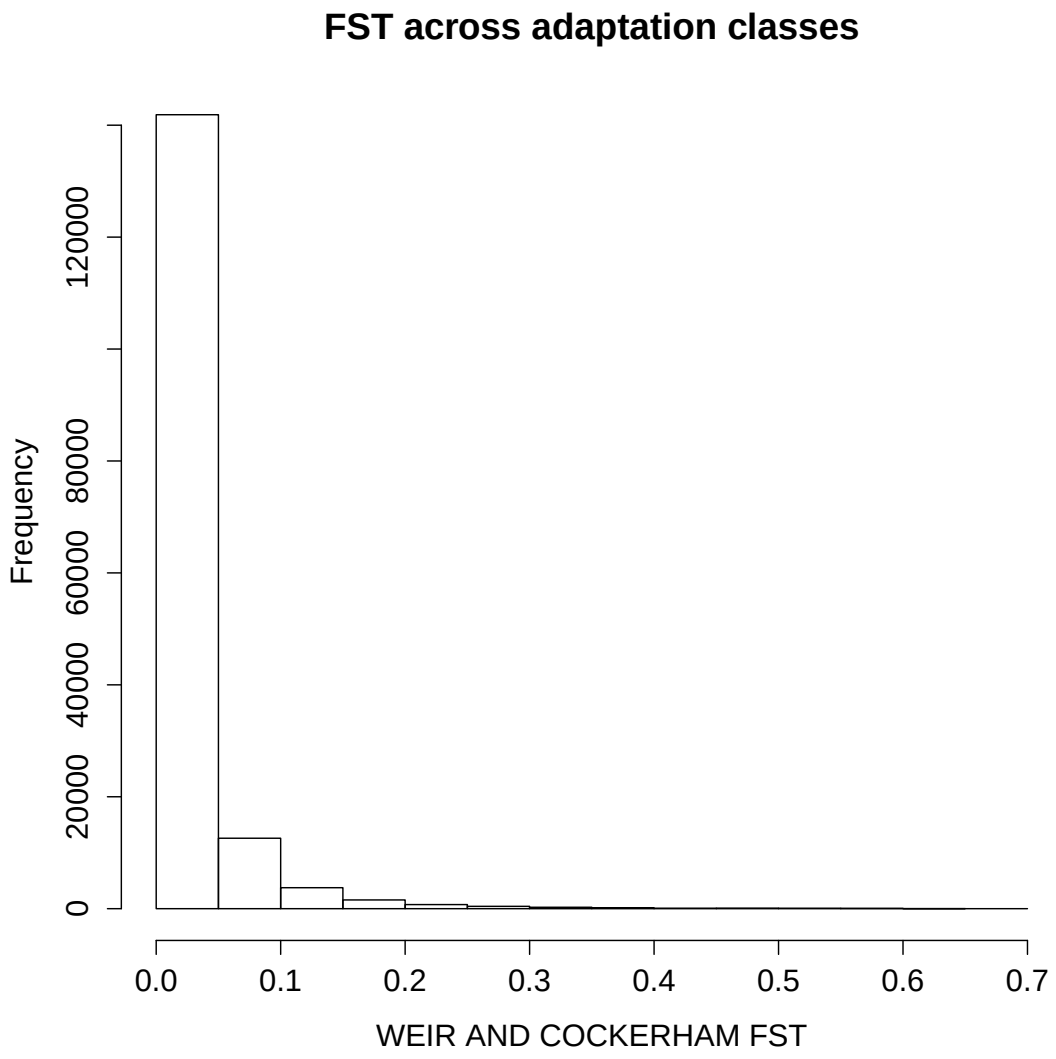


Figure 3.1: Distribution of genome-wide estimates of fixation index between adaptation classes

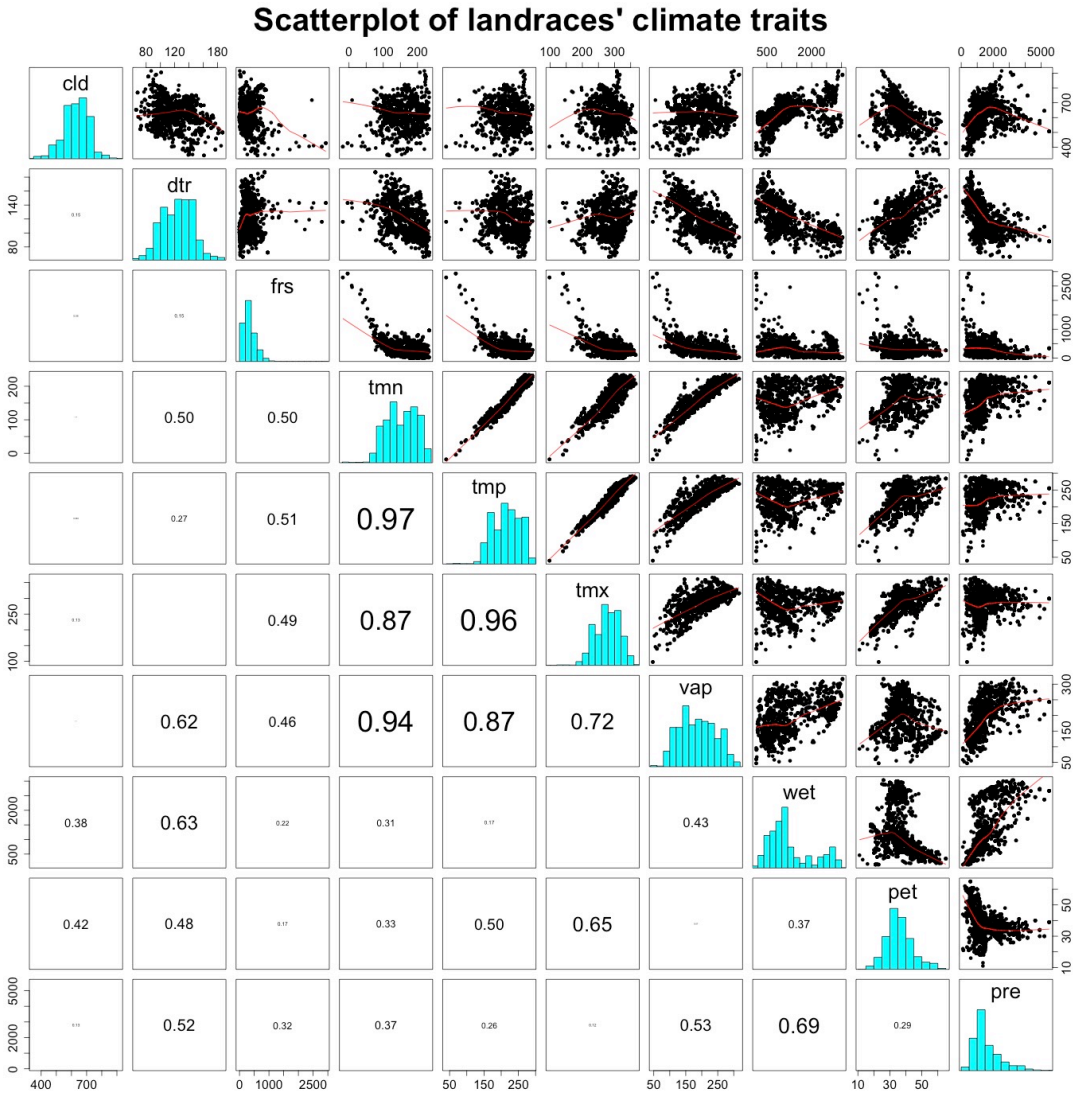


Figure 3.2: Distribution of climate traits

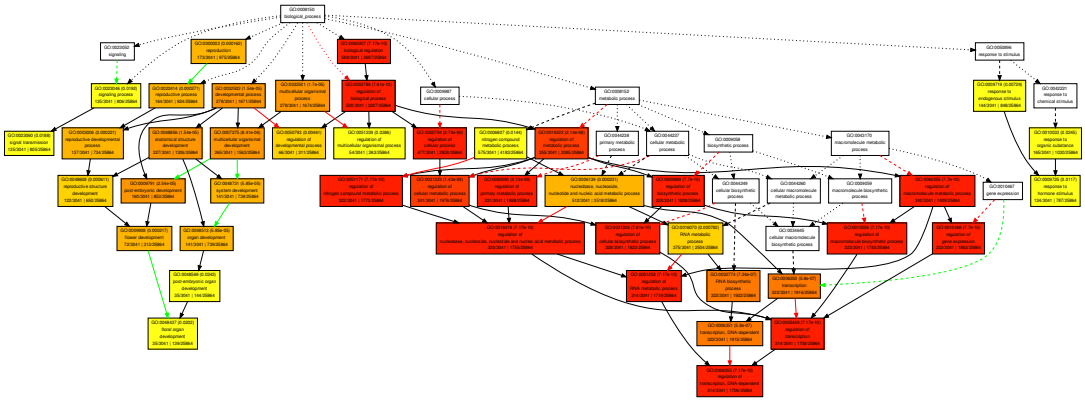


Figure 3.3: Gene Ontology enrichment for all climate associated genes

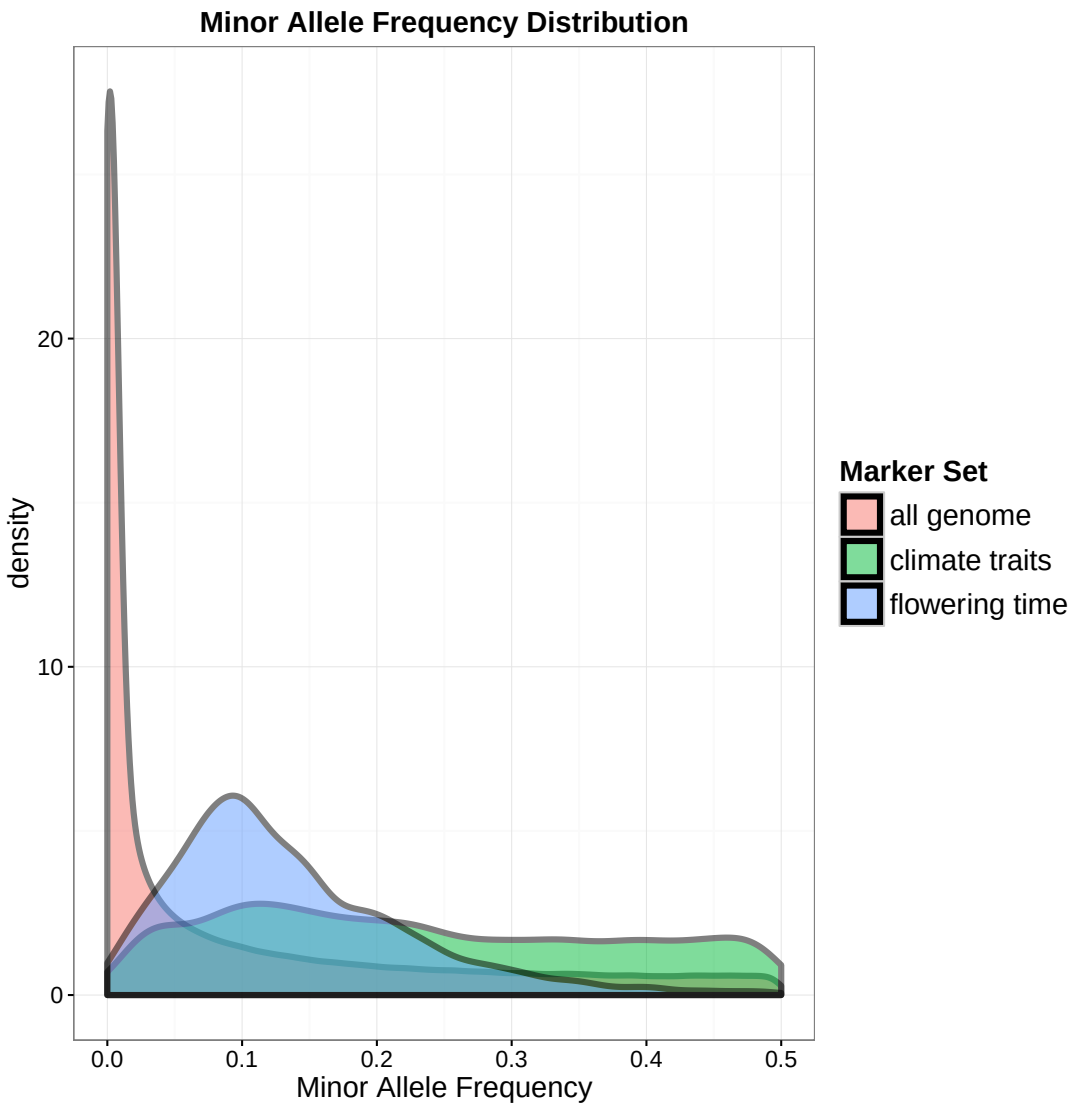


Figure 3.4: Minor allele frequency distribution for all, climate, and flowering time associated SNPs

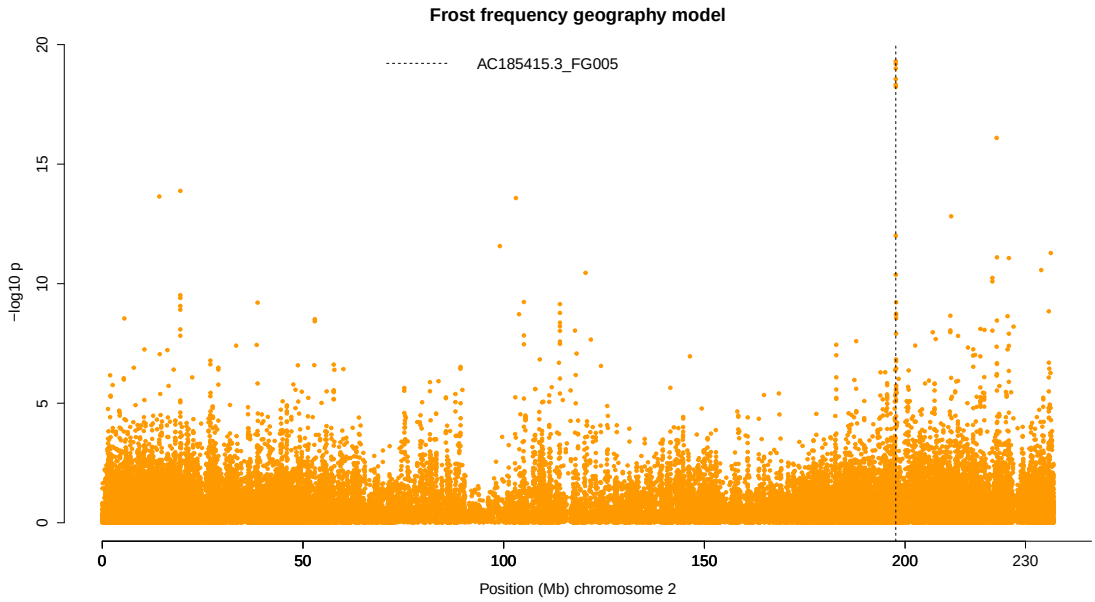


Figure 3.5: Chromosome with most significant hit for frost frequency

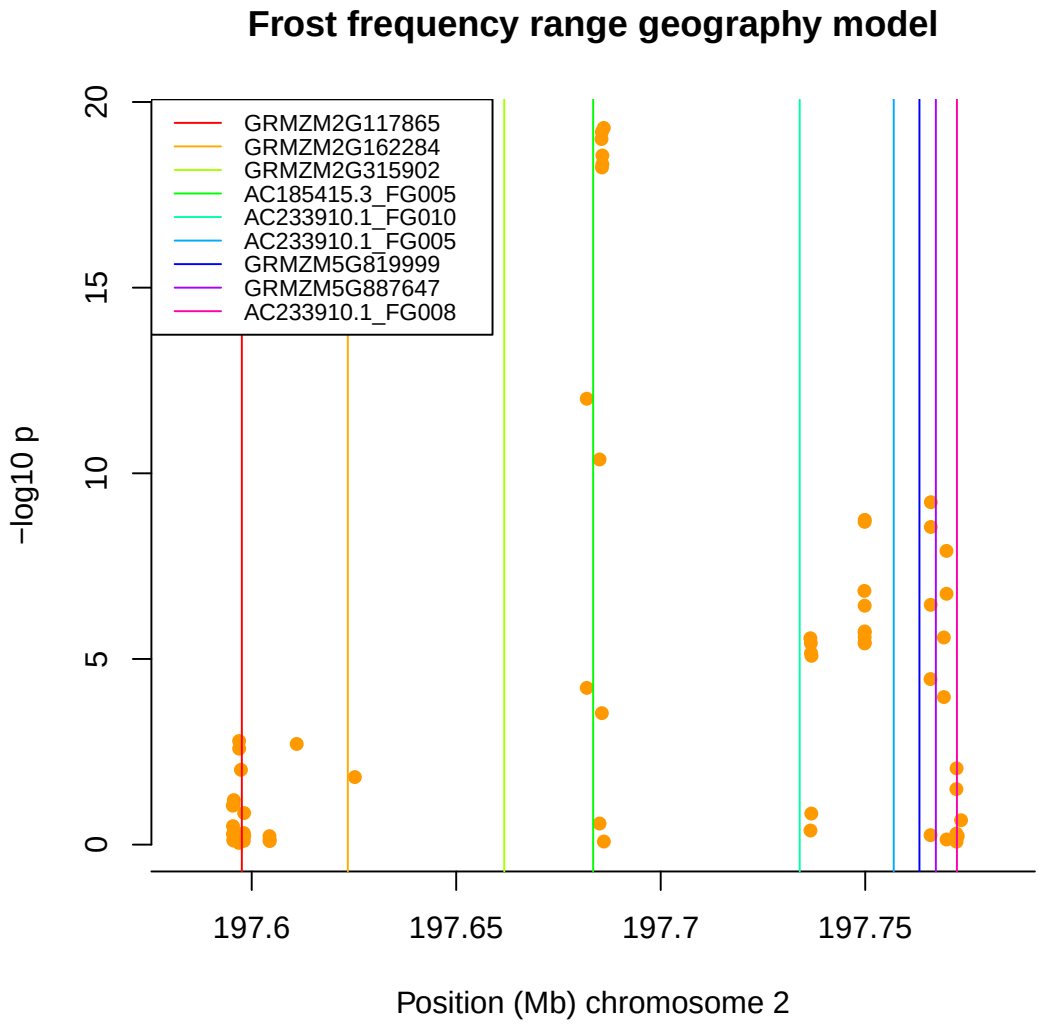


Figure 3.6: Genes around most significant hit for frost frequency

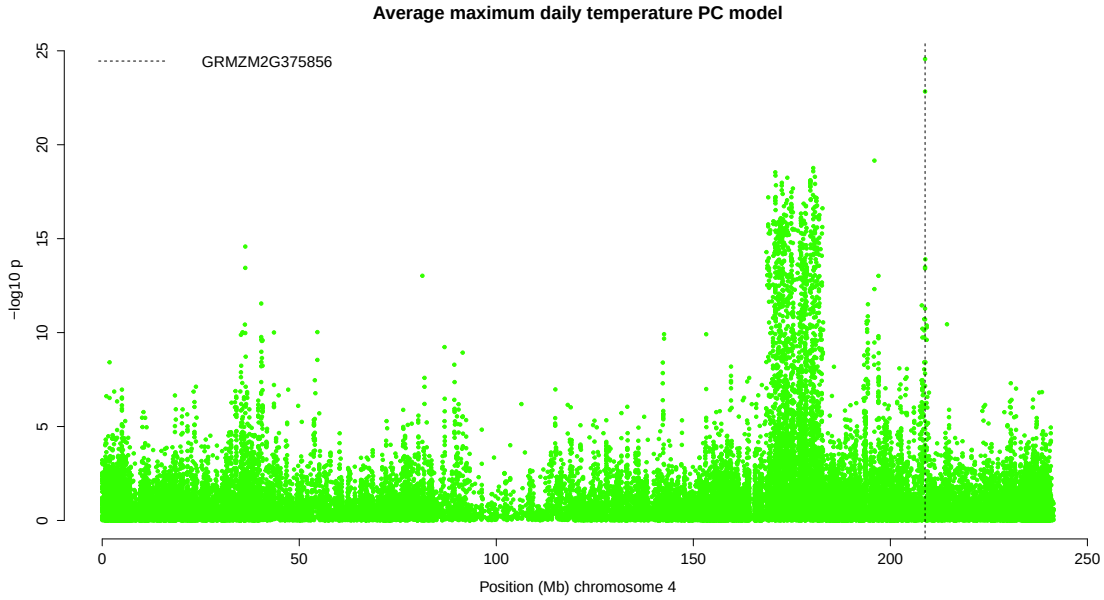


Figure 3.7: Chromosome with most significant hit for average maximum daily temperature

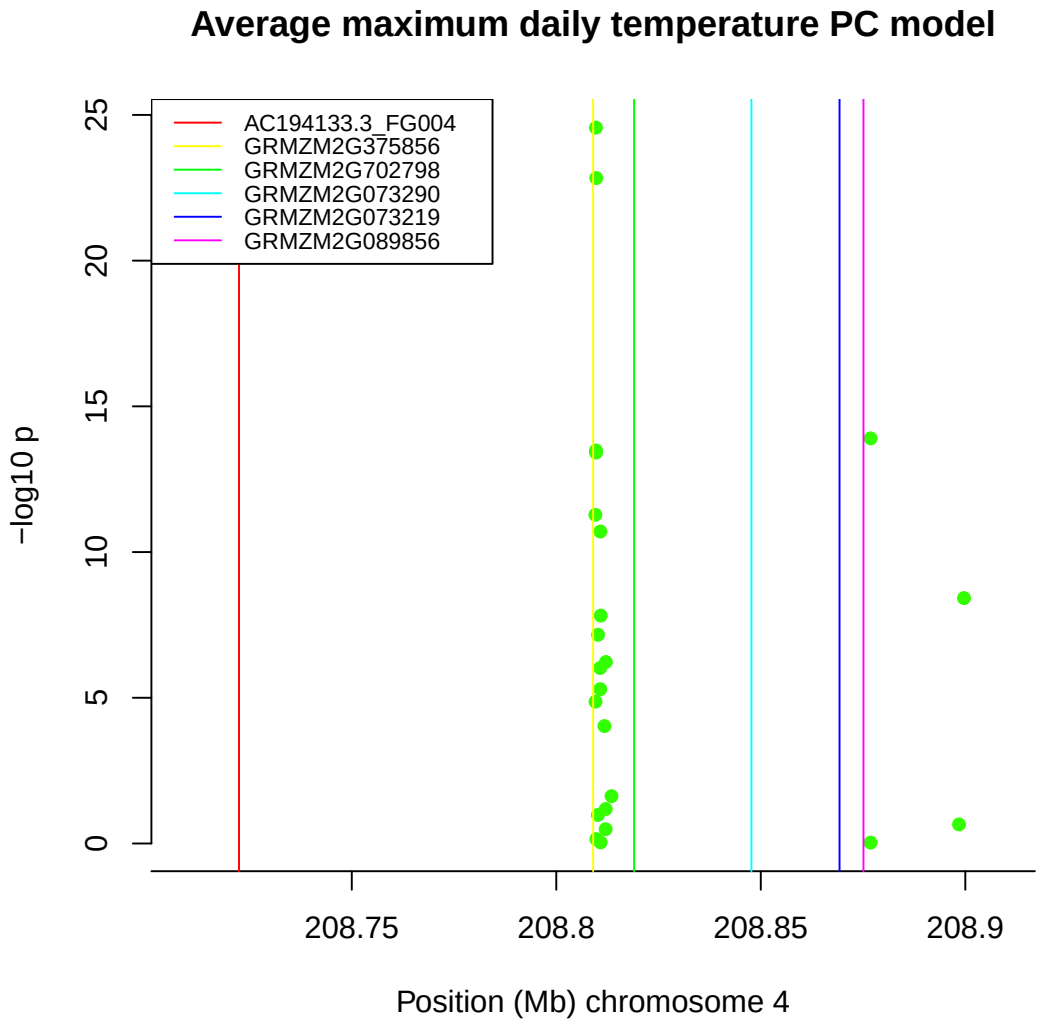


Figure 3.8: Genes around most significant hit for average maximum daily temperature

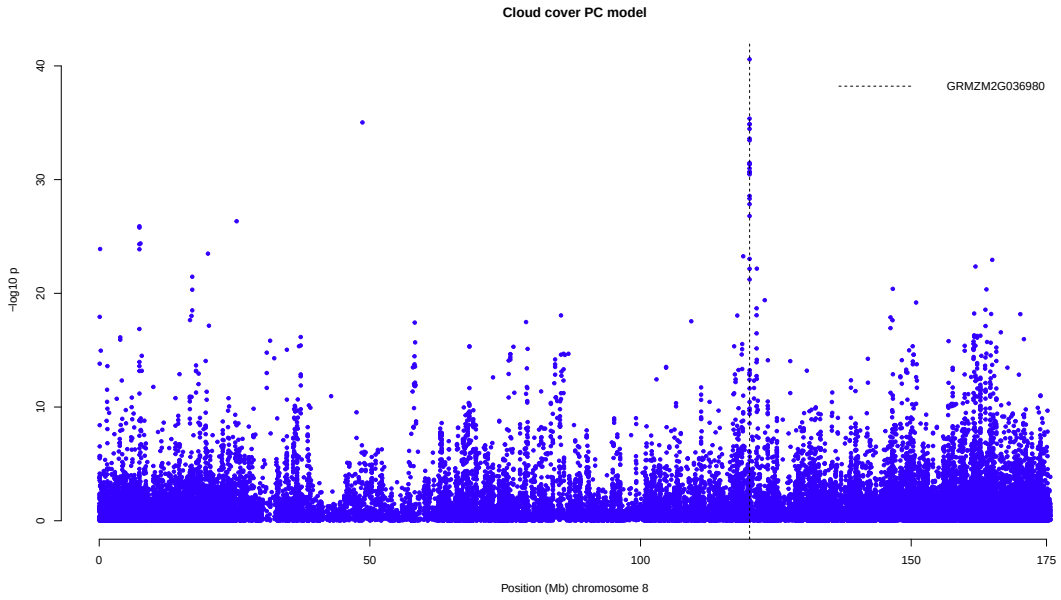


Figure 3.9: Chromosome with most significant hit for cloud cover

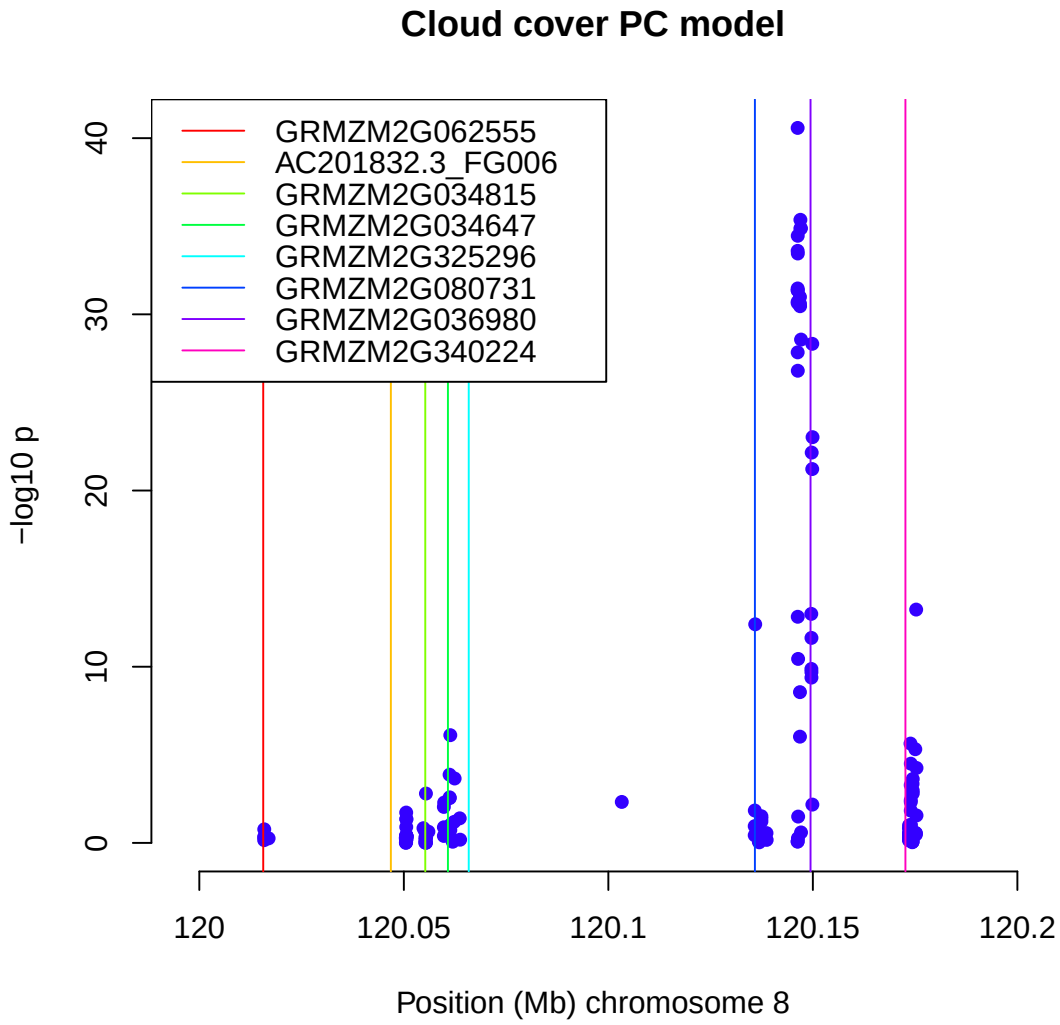


Figure 3.10: Genes around most significant hit for cloud cover

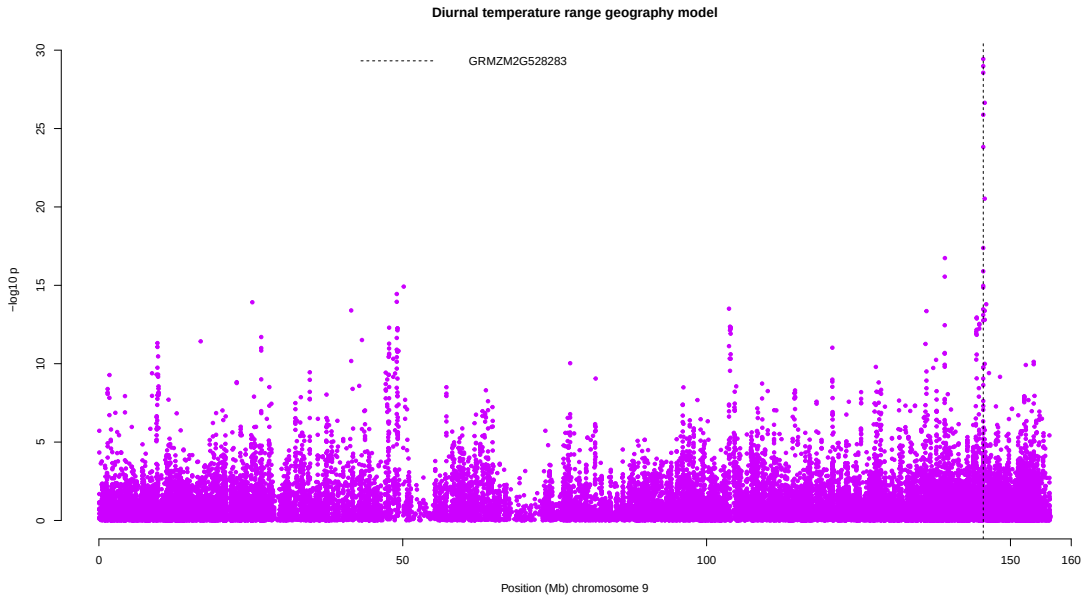


Figure 3.11: Chromosome with most significant hit for diurnal temperature range

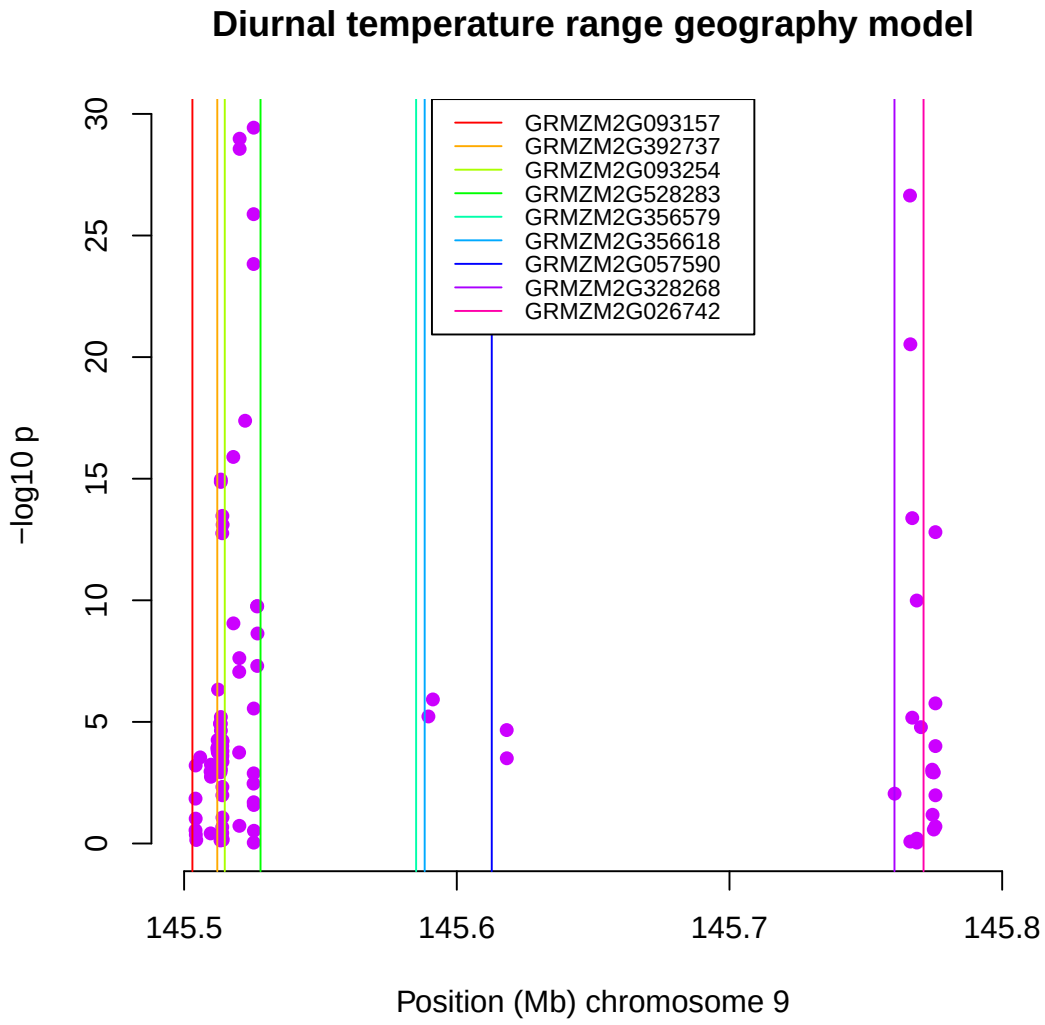


Figure 3.12: Genes around most significant hit for diurnal temperature range

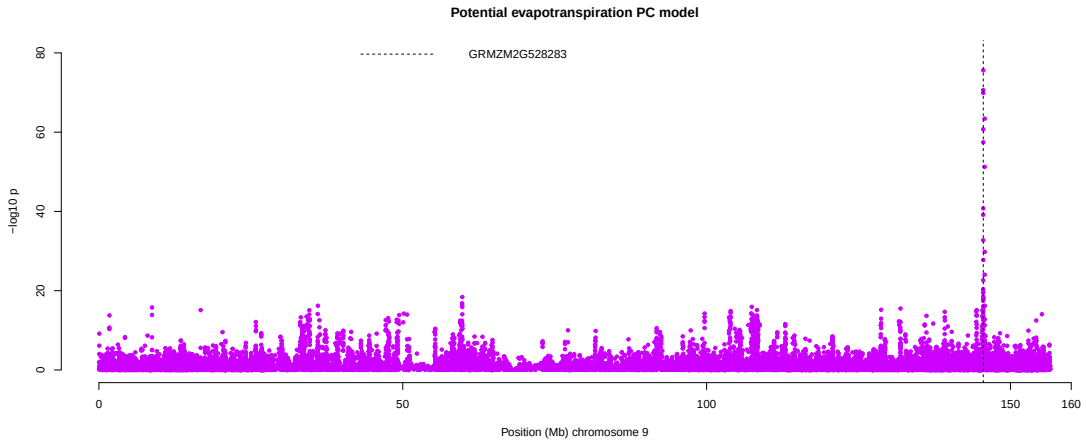


Figure 3.13: Chromosome with most significant hit for potential evapotranspiration

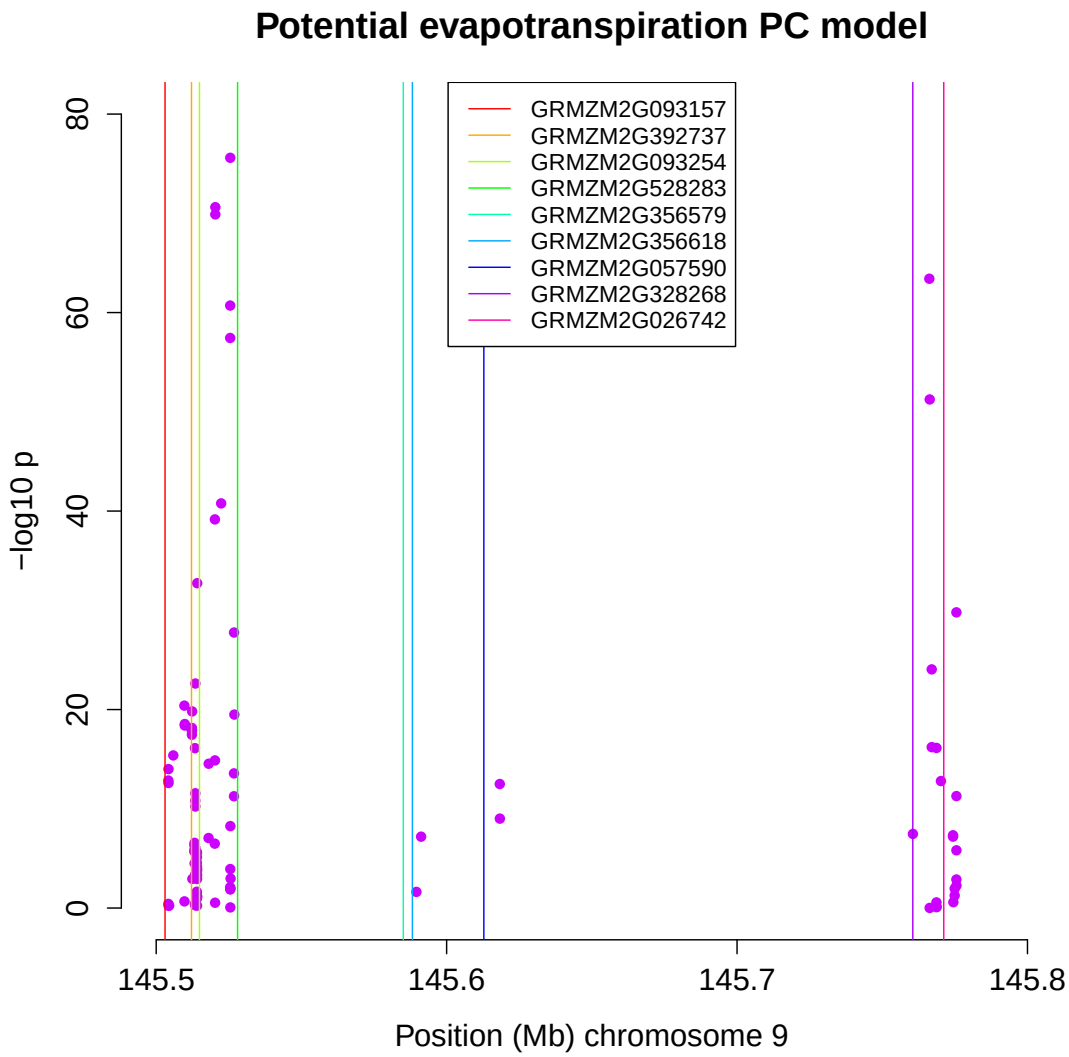


Figure 3.14: Genes around most significant hit for potential evapotranspiration

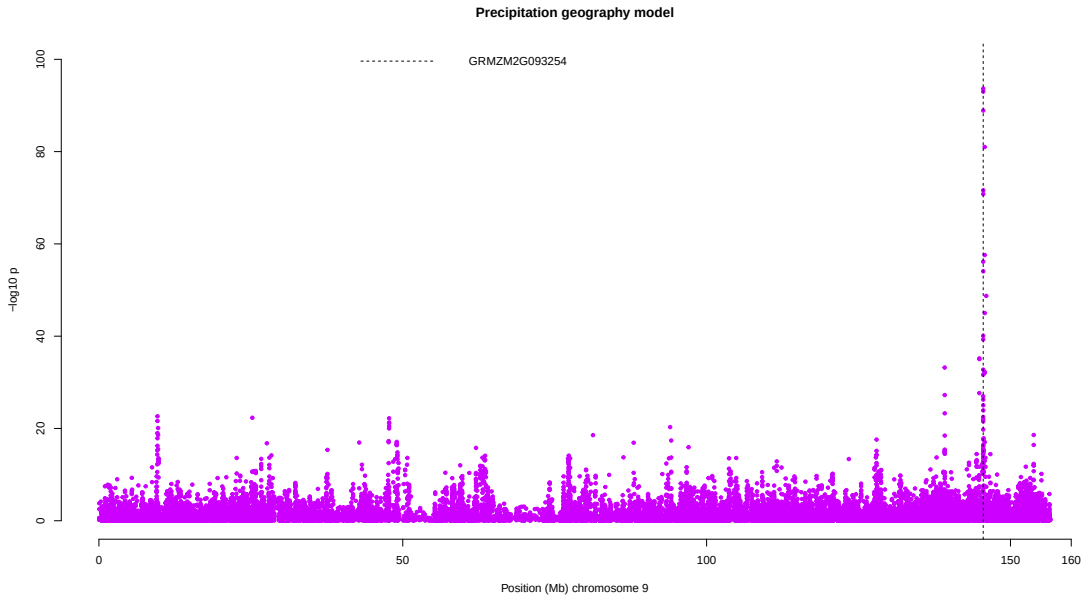


Figure 3.15: Chromosome with most significant hit for precipitation

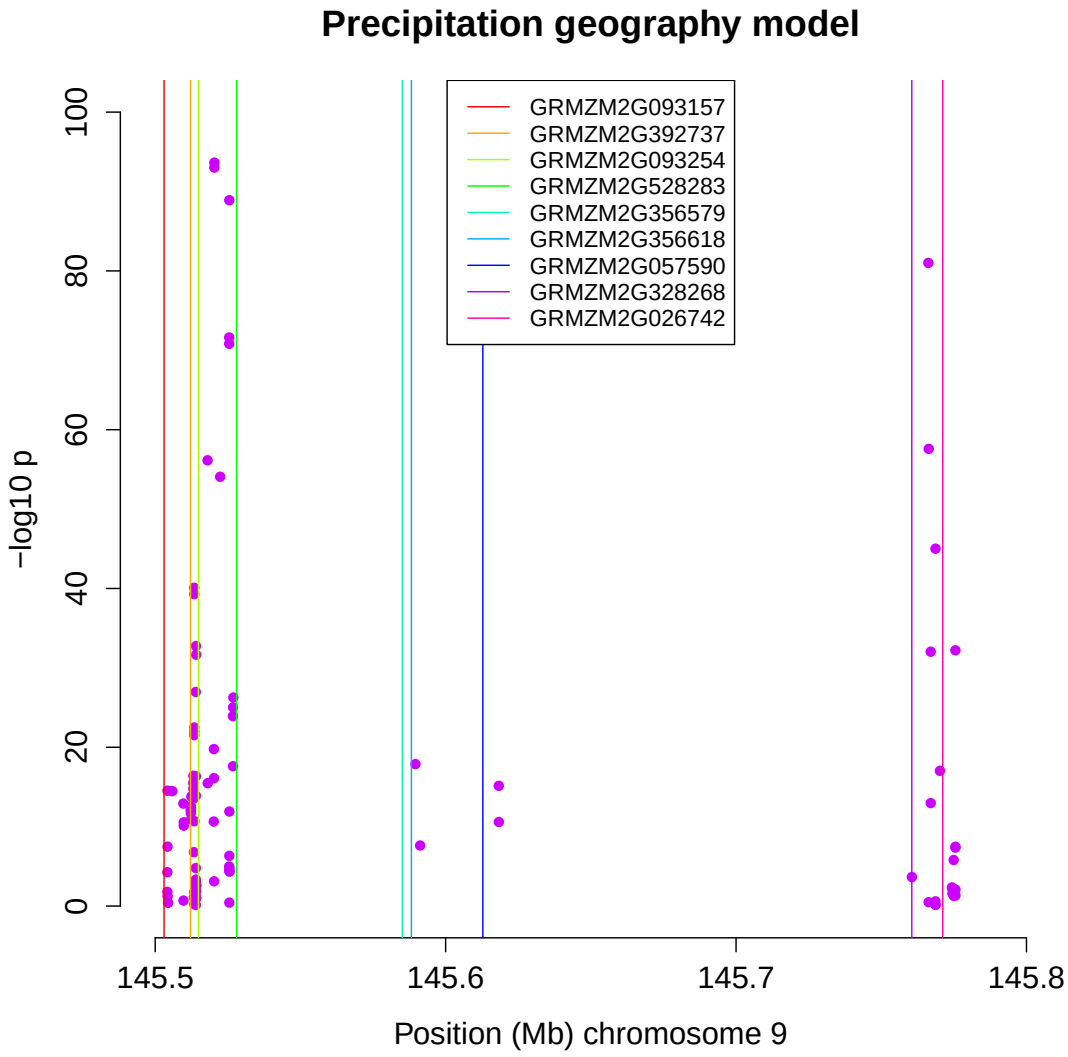


Figure 3.16: Genes around most significant hit for precipitation

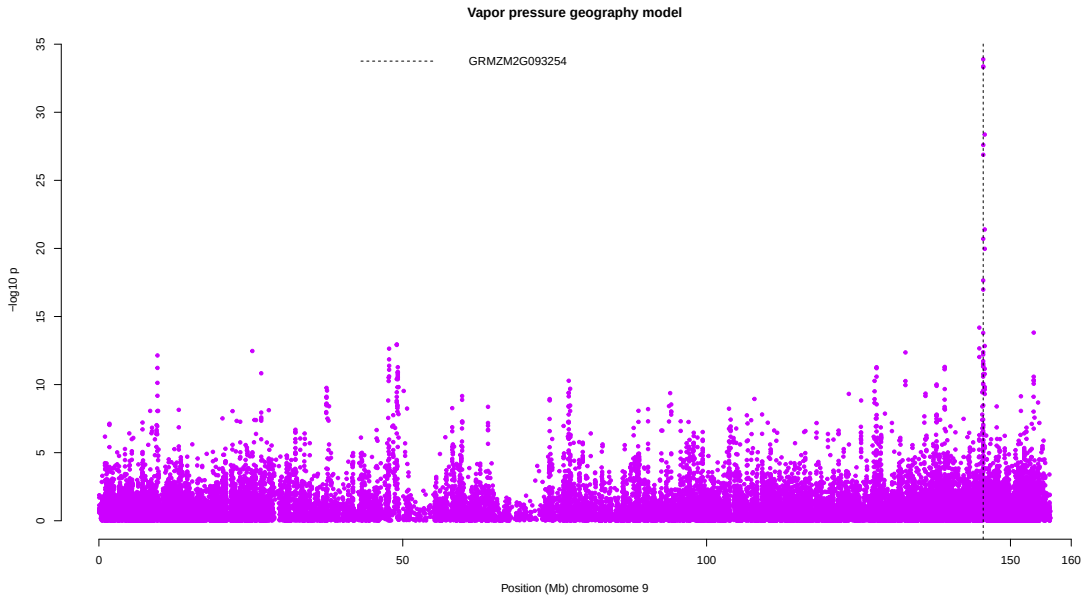


Figure 3.17: Chromosome with most significant hit for vapor pressure

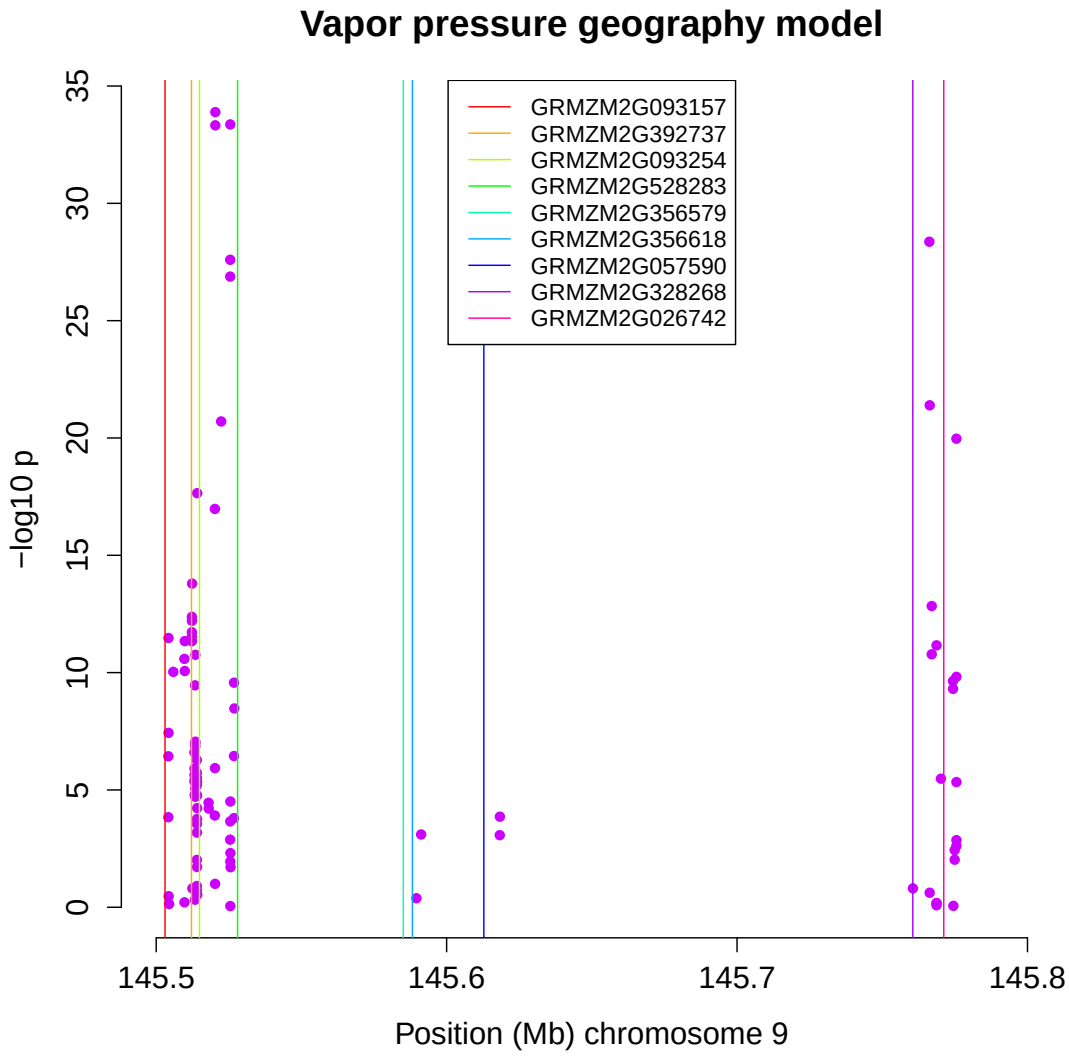


Figure 3.18: Genes around most significant hit for vapor pressure

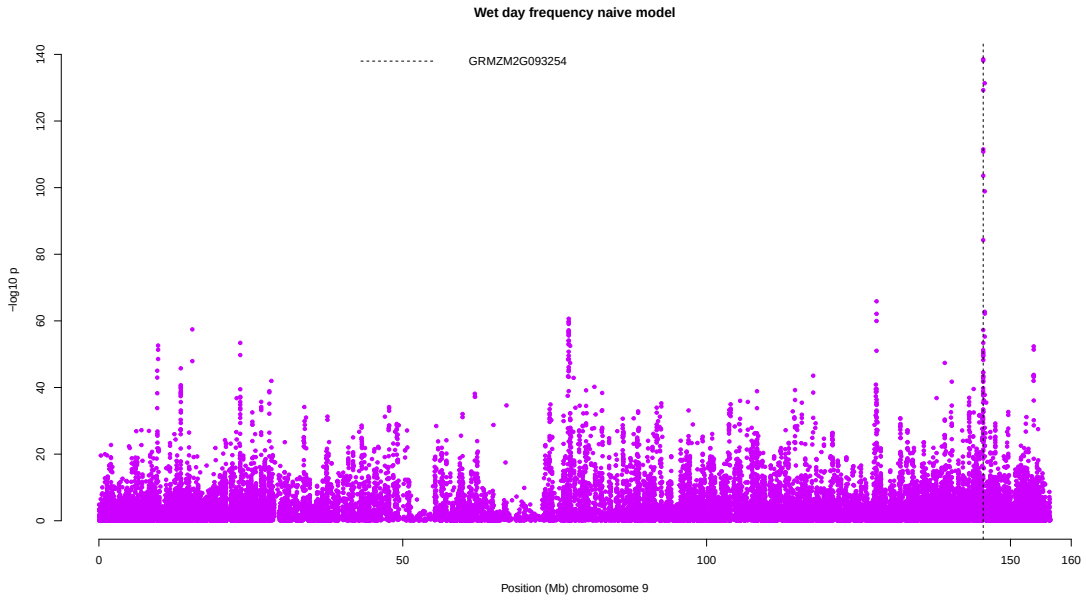


Figure 3.19: Chromosome with most significant hit for wet day frequency

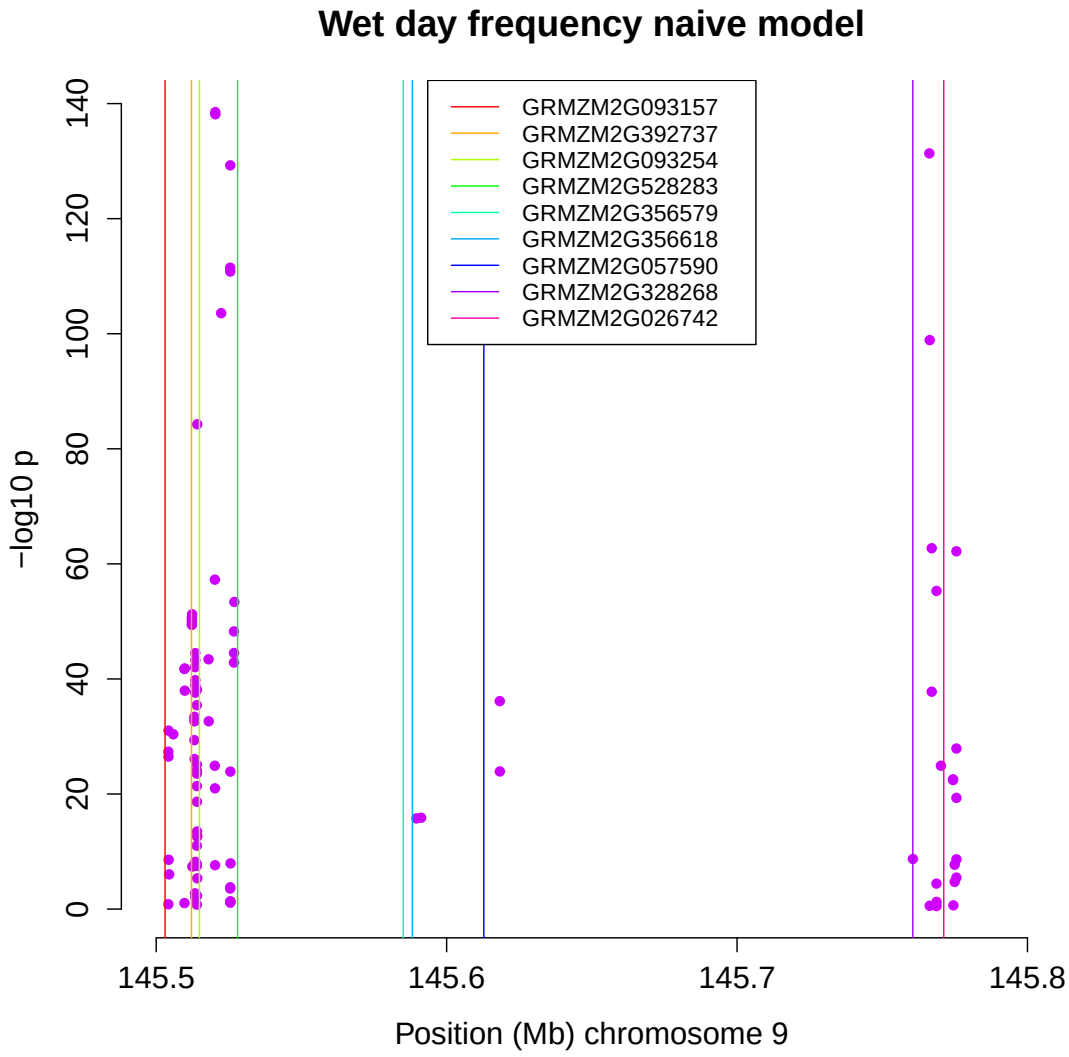


Figure 3.20: Genes around most significant hit for wet day frequency

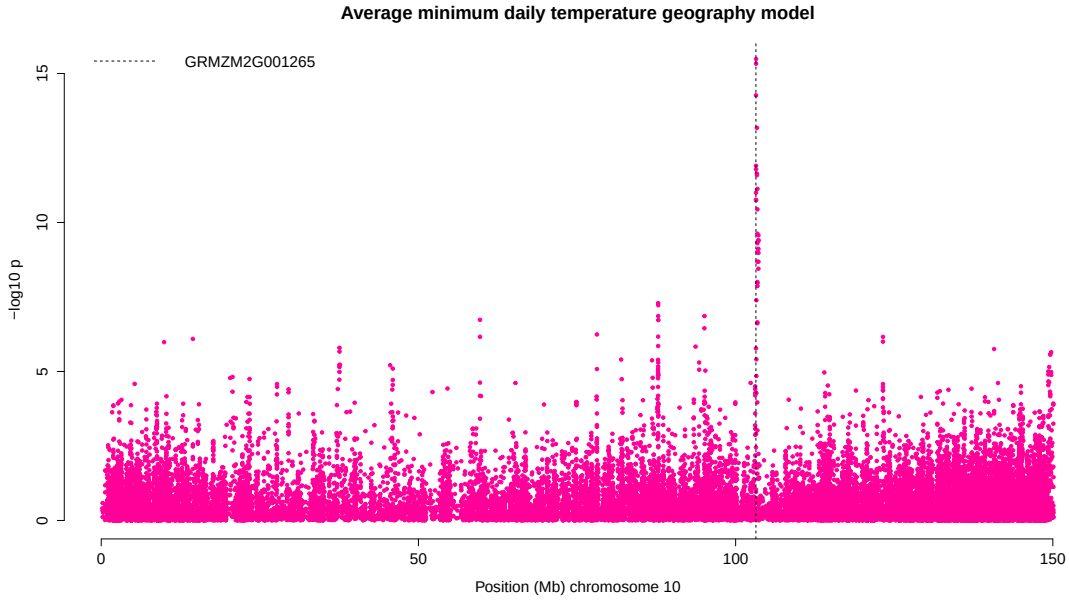


Figure 3.21: Chromosome with most significant hit for average minimum daily temperature

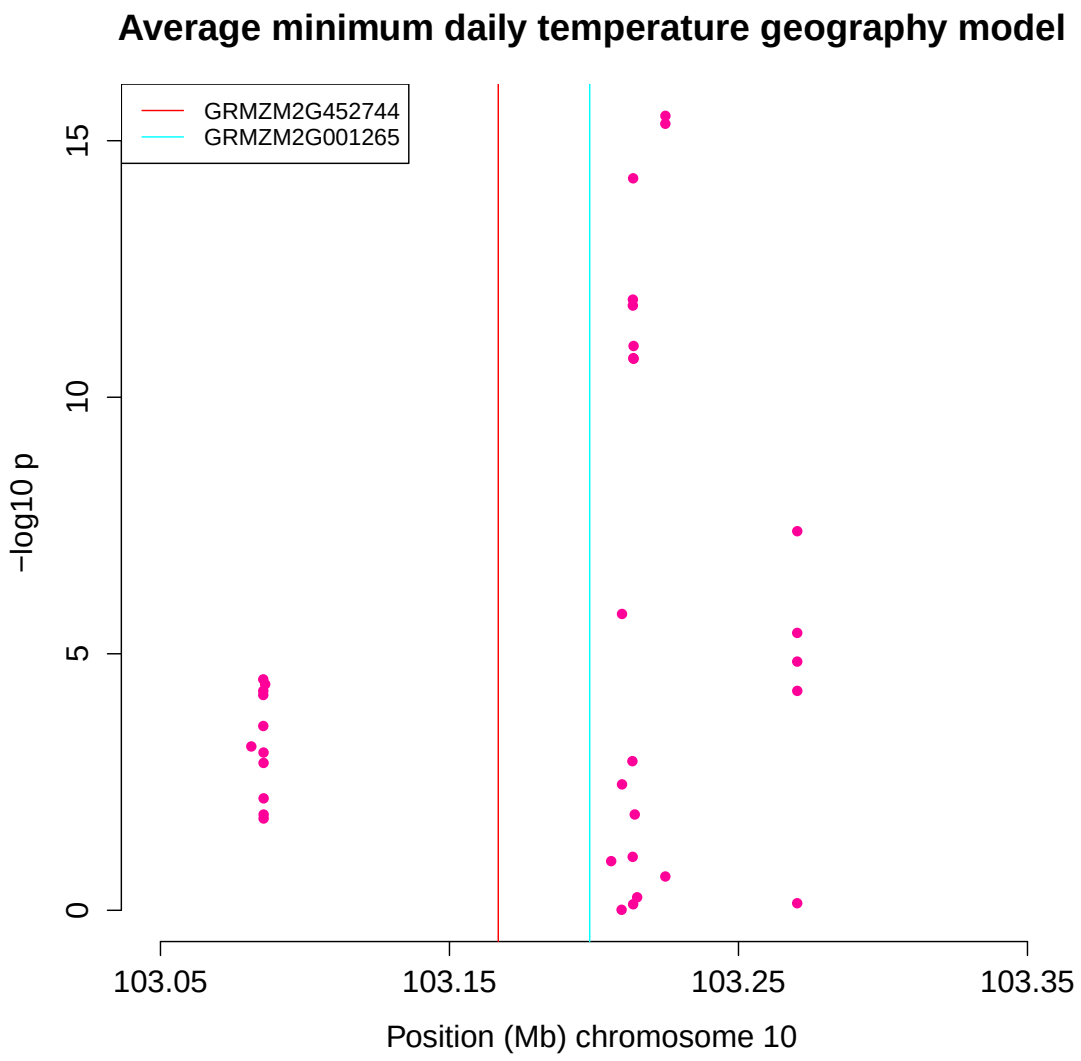


Figure 3.22: Genes around most significant hit for average minimum daily temperature

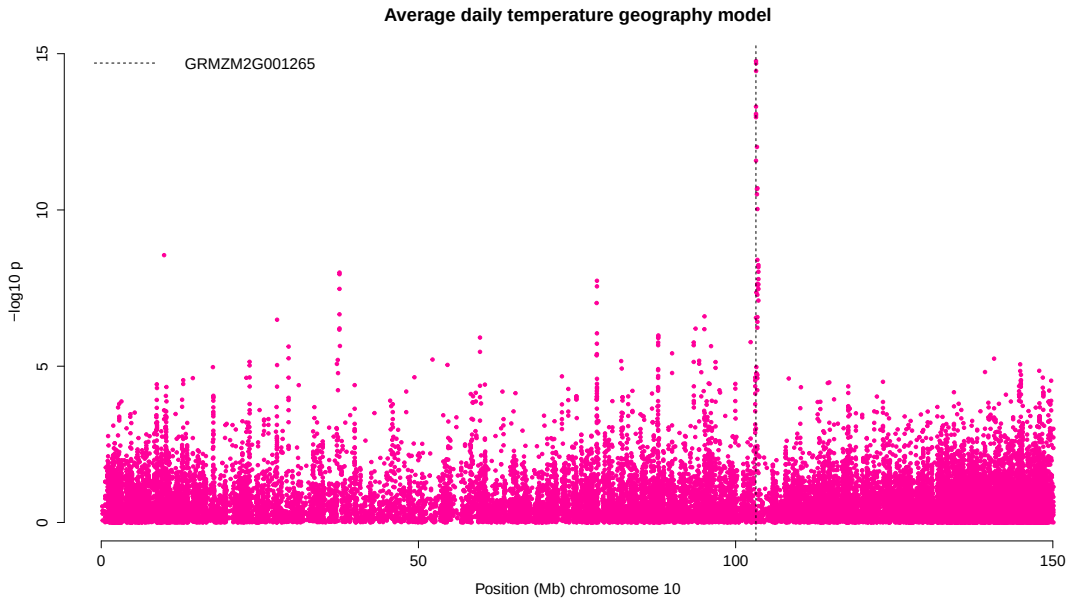


Figure 3.23: Chromosome with most significant hit for average daily temperature

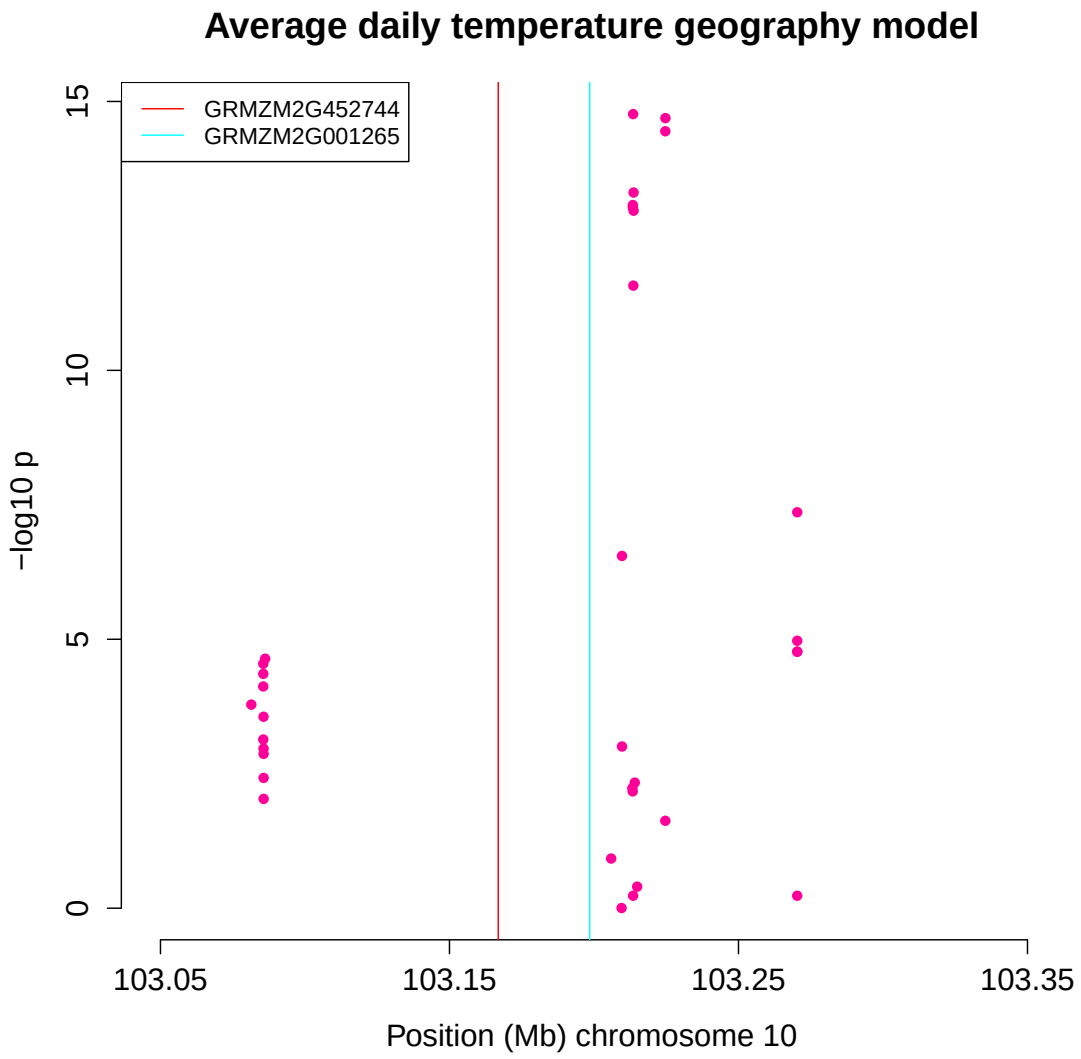


Figure 3.24: Genes around most significant hit for average daily temperature

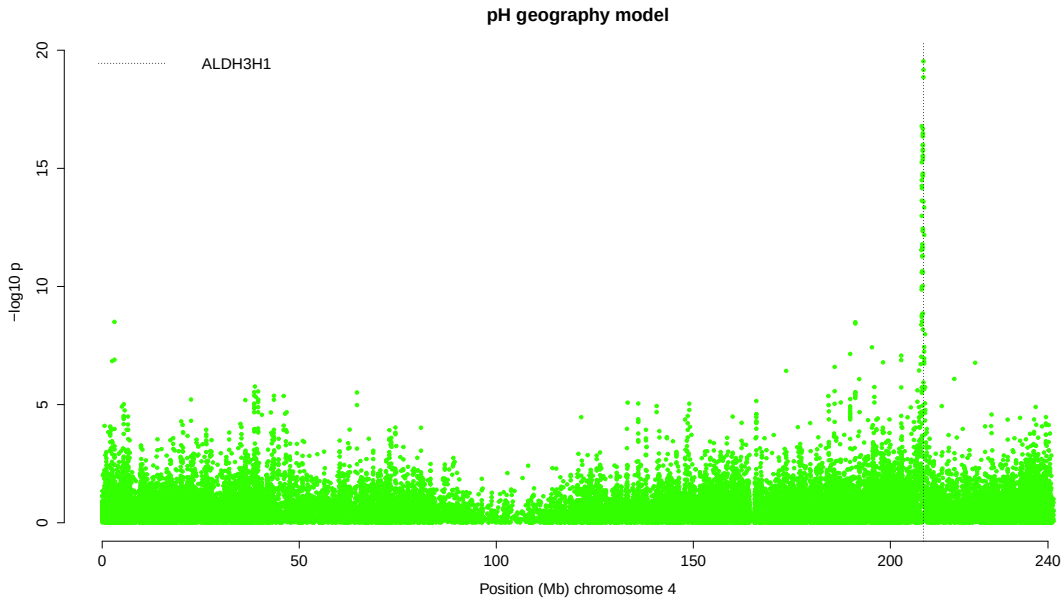


Figure 3.25: Chromosome with most significant hit for soil pH

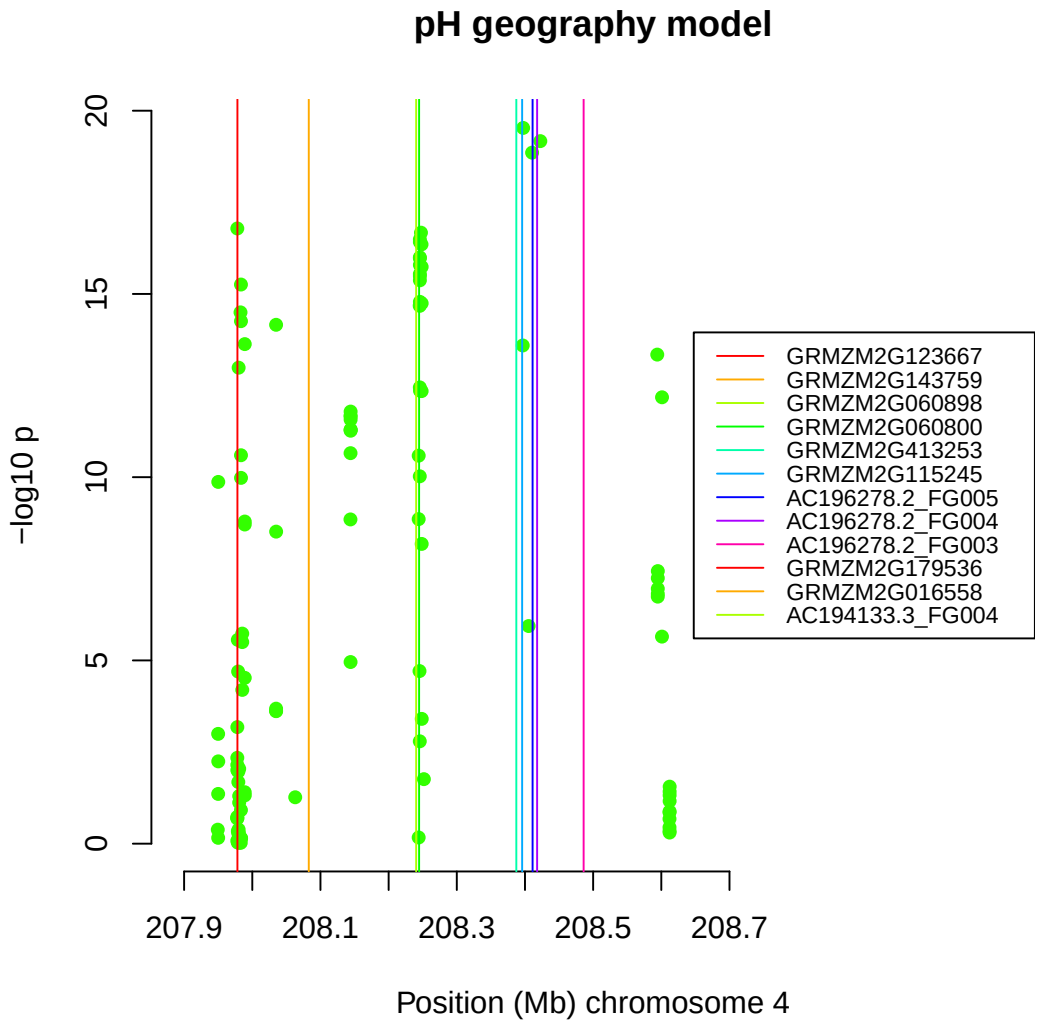


Figure 3.26: Genes around most significant hit for soil pH

3.6 Tables

Trait	Models	Proportion Structural	Proportion introgression	Significant Genes
cloud cover	naïve ∩ PC2	0.44	0.35	572
diurnal temperature rage	Geo ∩ PC2	0.45	0.35	465
frost frequency	naïve ∩ Geo	0.43	0.34	578
potential evapotranspiration	Geo ∩ PC2	0.43	0.33	550
precipitation	naïve ∩ Geo	0.07	0.02	1432
average minimum daily temperature	naïve ∩ Geo	0.65	0.55	225
average daily temperature	naïve ∩ Geo	0.78	0.68	122
average maximum daily temperature	Geo ∩ PC2	0.67	0.59	201
vapor pressure	naïve ∩ Geo	0.21	0.12	296
wet day frequency	naïve	0.12	0.08	3046

Table 3.1: Climate top models, number of genes, and contribution of high-LD and introgression

Gene	*Arabidopsis* homolog	Annotation in *Arabidopsis*
GRMZM2G439422	AT5G47000.1	Peroxidase superfamily protein
GRMZM2G458548	AT5G38560.1	Encodes a member of the proline-rich extensin-like receptor kinase (PERK) family.
AC190909.3_FG004	AT5G13000.1	Encodes a gene similar to callose synthase
GRMZM2G059212	AT5G13000.1	Encodes a gene similar to callose synthase
GRMZM2G396825 (glk19)	AT4G16110.1	Encodes a pollen-specific transcription factor
GRMZM2G328268	AT3G55060.1	unknown protein
GRMZM2G151496	AT2G24540.1	ATTENUATED FAR-RED RESPONSE (AFR)
GRMZM2G048472	AT2G18950.1	Encodes homogentisate phytyltransferase involved in tocopherol biosynthesis.
GRMZM2G093254	AT2G01110.1	Core subunit of the chloroplast Tat translocase.
GRMZM2G045171	AT1G73370.1	Encodes a protein with sucrose synthase activity (SUS6)
GRMZM2G469828	AT1G70090.1	GATL9 Encodes a protein with putative galacturonosyltransferase activity.
GRMZM2G361693 (pdh1)	AT1G59900.1	encodes the e1 alpha subunit of the pyruvate dehydrogenase complex (PDC)
GRMZM2G141222	AT1G28540.1	unknown protein
GRMZM5G885045	AT1G20823.1	Encodes a RING E3 ubiquitin ligase ATL80

Table 3.2: Top genes associated to multiple climate traits

CHAPTER 4

GENOME WIDE ASSOCIATION FOR PLANT HEIGHT VARIATION IN MAIZE LANDRACES

4.1 Abstract

Plant height is a key agroecological trait. For most plants, height is an adaptive trait involved in competition, with increased elongation arising from low-light conditions. In addition, since domestication, plant height has been modified for several species, and in particular the use of dwarfing phenotypes was fundamental to the recent rise in yield from the Green Revolution. Similar to human stature, plant height is highly heritable, and in both plants and humans the genes underlying height variation in natural populations remain elusive. Here, we explored the genetic basis of plant height variation in a comprehensive panel of 4,100 landraces from Latin America. We compiled a list of maize homologs to *Arabidopsis* genes involved in hormonal pathways, and significant enrichment was observed for the genes associated with plant height in this panel. We further showed significance with one domestication gene, *grassy tillers1*, which is involved in plant architecture in maize. We observed sharing with significant genes previously associated with height in maize inbred lines as well genes associated with flowering time in the same landrace panel. By comparing the allele frequencies with SNPs associated flowering time, we observe a significant shift towards low frequency polymorphisms, suggesting that rare and likely deleterious alleles are enriched in the genetic control of this trait. We perform genome wide prediction and show that plant height remains with limited pre-

[0]For this chapter, my contribution encompassed analyses using the genotypic data, the genome wide association model and subsequent gene level analyses

dictive ability at 500,000 markers. Together this results provide new candidates for understanding the genetic control of plant height in natural populations.

4.2 Introduction

Plant height is a highly heritable trait that has been subject of study in quantitative genetics for over a century. From an evolutionary perspective, across plant species, differences in height are under strong selection. Plant height is a critical trait for plant-plant competition, with increased elongation in response to changes in light composition. This physiological response is controlled by phytochromes through what is known as the shade avoidance response syndrome[1], and it mediates competition between closely located plants. In terms of selection for differences in agronomic practices, plant height changes in the form of dwarf varieties became a key player in the Green Revolution, allowing significant increases in yield for crops like wheat[2], and rice[3].

In addition to their use in agriculture, the study of dwarf mutants has helped elucidate several enzymes involved at key steps in the biosynthetic and regulatory pathways of different hormones. The genes behind dwarfing effects are generally single genes with big effects affecting key steps in plants hormonal pathways[4] including in maize gibberellin acid[5] (GA), auxin [6], in *Arabidopsis* brassinosteroids[7] and in rice dwarf mutants there is evidence of co-regulation of plant stature and tiller numbers[8]. Mutants in maize affecting gibberellin biosynthesis include *Anther ear 1* (*An1*), with affected plants displaying typically semi-dwarf phenotype. The An1 gene product is responsible for the cyclization of geranylgeranyl pyrophosphate to ent-kaurene, the first step

in the biosynthesis of gibberellins[9]. The enzyme that catalyses this step is also likely redundant in the maize genome. Bioassay experiments further show that the four additional maize mutants *dwarf-1*, *dwarf-2*, *dwarf-3*, and *dwarf-5* contain each mutations that affect different steps of the GA biosynthesis pathway[10,11]. Genes downstream of GA biosynthesis involved in signaling can also lead to dwarfing, including in maize, the mutants *dwarf plant8* (d8) and *dwarf plant9* (d9), which encode DELLA proteins[12]. DELLA proteins are negative regulators of GA action that are degraded following their interaction with the GID (GA-INSENSITIVE DWARF1)-GA complex[13]. Auxins are another class of plant hormones with significant role in many aspects of plant development. Unlike gibberellins, the biochemical pathways for auxin biosynthesis and downstream signalling are more complex[14]. Auxin plays a central role in cell expansion and division, and in maize the *brachytic2* (*br2*) mutant displays dwarfism due to a loss of function mutation that modulates polar auxin transport in the maize stalk[6]. In *Arabidopsis*, there is also evidence of interaction between auxins and brassinosteroids, a group of hormones involved in plant growth and development, affecting lateral growth[15]. Mutants in the production or perception of brassinosteroids display dwarfism in *Arabidopsis* and have been used to characterize the underlying pathways[7]. In maize, the *nana plant1* (*na1*) mutant displays significantly reduced height as well a feminized male flowers, and contain a mutation in a gene homologous to DET2, which is involved in brassinosteroids biosynthesis[16]. Furthermore, transgenic maize plants for ZmDWF1 display varying levels of dwarfism. ZmDWF1 is the maize homolog of the Arabidopsis DWF1/DIM gene, which in that organism is involved in brassosteroids biosynthesis. In many grasses, height also displays a strong correlation with flowering time[17], however for natural standing varia-

tion the genetic basis for both traits is not entirely shared[18].

Since the availability of low-cost high-density genetic marker platforms, genome wide association as well as random effect models have been used to study the genetic architecture of quantitative variation in stature in several species. In the case of human height, random effect models including all common variants can capture 60% of the 80% trait heritability, however significant genes from genome wide association each explain only a small fraction of the phenotypic variance [19]. In maize, plant height has been studied in inbred lines, showing that it is also a highly heritable yet complex trait that can generally be predicted accurately from relatedness using random effect models, however the genetic basis for stature variation can be best explained through an infinitesimal model, with numerous genes each with very small effects [18]. Although significant markers associated with plant height have been identified in maize inbred lines, association between stature and natural standing variation in landraces, the largest reservoir of maize genetic diversity, remains uncharacterized.

We explored the genetic architecture of plant height in a diverse panel of maize landraces. We used the progeny and genotypic data from the F-One Association Mapping population[20], which represents the diversity of 36 countries in Latin America. We performed evaluation in 30 trials across Mexico, with 6 trials under Nitrogen or water stress. Using Genome Wide Association, we found 1056 genes to be associated with plant height. We compared our associating genes to a those of previous mapping populations in maize, and to a list of genes homologous to those involved in hormonal pathways in *Arabidopsis*. Genomic selection was also performed, and prediction accuracies are compared

across marker densities.

4.3 Results

Phenotypic evaluation and estimation of breeding values

The main goal of the project was to characterize the genetic basis of height variation across maize landrace accessions. In total 3,425 F1 progeny accessions from the maize F-One Association Mapping (FOAM) population[20] were used to set up field trials. The two main limiting factors for setting up field evaluation were space and adaptation to elevation of the landrace accession. To avoid potential confounding effects of lack of adaptation, following the original design accessions were tested only on locations of matching adaptation, and the number of accessions per location was determined by field space. At each location, experimental rows were set up with 10-15 plants per accession in 30 locations across Mexico (Methods, Table 4.8). The number of accessions evaluated on each trial ranged from 282 in the smallest trial to 1924 on the largest, with a mean of 821 accessions per location. Plant height was measured on 5 plants per accession per location.

After collection of field measurements, phenotypic data was analysed using a mixed linear models in order to get fitted means for each accession across all trials. Models included main fixed effects for hybrid parent, a random accession effect and the random autoregressive effect to account for field variation. Across all trials for all landrace accession, a total of 24,744 height adjusted values were estimated, with values varying across all locations from 101.3 cm to 364.1 cm

with a mean of 246.7 cm. In maize populations, as well is some other grasses, high correlation between flowering time and plant height have been reported [17,18,21]. In order to estimate the extent of phenotypic correlation between both traits in the landrace panel, fitted height estimates were compared with fitted flowering time estimates at the 23 trials with data available for both traits. High correlation between flowering time and plant height was observed (Supplementary figure 4.5), with median correlation value across all trials of 0.63. Only one location (m2012bal), displayed no correlation.

Lists of association candidates

Height variation is a complex trait, and there are different lines of evidence from various studies regarding potential genes underlying its control. Four lists of candidate genes were compiled to test for overlap after performing genome wide association. The first list contained the significant markers associated with plant height on previous mapping experiments using inbred lines[18]. Those results include a) the 3,328 hapmapv1 and hapmapv2 markers identified as significantly associated with plant height in the NAM population, projected onto AGPv2 coordinates, as well as b) the 3,779 markers significantly associated with plant height in the NCRPIS panel of inbred lines[18]. The second list was assembled in order to identify potential candidate genes involved in hormonal biosynthesis or regulation. The list included the maize homologs of the 650 unique genes annotated in *Arabidopsis*[22] as having genetic or transgenic evidence for effect in the following 8 hormonal pathways: abscisic acid, auxin, brassinosteroid, cytokinin, ethylene, gibberellin, jasmonic acid, and salicylic acid. In total 1045 genes were identified as homologs to the hormone related

genes in *Arabidopsis*. From the total, 85% of the genes were annotated within a single hormonal pathway, and the remaining 15% corresponded to two or more hormones. In terms of their role within hormonal pathways, the largest class containing 67% of the genes corresponded to annotations in hormone signal transduction, followed by 11% involved in hormone biosynthesis, 10% hormone response, 5% as hormone receptor, and 2% in hormone transportation and metabolism. The third list was compiled to explore the extent of genetic overlap between flowering time and plant height, and contained respectively the 881 and 883 genes significantly associated with female and male flowering time on the same panel[20]. The fourth and final list contained all 39,423 protein coding genes annotated in the maize genome filtered gene set, and was used in order to characterize previously unknown genes involved in height variation in maize landraces.

Genome wide association

Genome wide association was performed for all trials combined using a two step mixed linear model accounting for location, population structure and relatedness, with SNP effect test nested within trial (Methods). Based on the distribution of statistical significance across all sites, only the top 0.5% of the markers (n=2,467) based on p-value (-log10(p)>9) were considered as significantly associated with plant height variation (Figure 4.1, Supplementary figure 4.5). The previous study of flowering time variation on the FOAM landrace panel reports enrichment of significant markers at regions with increased LD and/or Zea mays ssp. mexicana introgressions[23]. In contrast, for plant height a modest contribution of introgressions and high-LD regions was observed, with a

proportion of only 0.05 and 0.14 respectively. Unlike flowering time, the large inversion on chromosome 4 was not significantly associated with plant height, however the putative inversion on chromosome 3 and the centromeres of chromosomes 3 and 5 were both associated with height as well as flowering time. Outside the low recombination regions, we observed significant association of 1056 genes with plant height. Those significant genes encompassed 40 significantly enriched ontology terms with the most significant corresponding to anatomical structure development (FDR p value < 0.05, Supplementary table 4.4).

The significant markers from the FOAM GWA were first compared to the list of candidate genes from previous height association results in other maize mapping populations. In total 122 markers representing 95 genes overlapped between the significant association results for the FOAM and NAM populations (Supplementary table 4.5), however no significant ontology enrichment was observed for the set of overlapping genes. For the NCRPIS panel, a total of 63 markers representing 41 genes overlapped with the markers significant FOAM results for plant height (Supplementary table 4)), with significant GO enrichment for terms related to ion binding (FDR p value ¡0.05). Interestingly, two major flowering time loci, ZmRap2.7 and ZCN8, the maize florigen and homolog to *Arabidopsis* FT[24] were significantly associated with height variation across all three maize panels. Furthermore, given the observed correlation between flowering time and plant height across most trials, it was expected to observe overlap between significant markers from GWA for both traits. A significant 3-fold enrichment was observed of flowering time associated markers among the significant height associated SNPs. Among the flowering related genes was also the maize gene GRMZM5G844173 (gi2 - gigantea2), homolo-

gous to *Arabidopsis* GI/ GIGANTEA where together with CONSTANTS (CO) and FLOWERING LOCUS T (FT), GIGANTEA promotes flowering under long days in a circadian clock-controlled flowering pathway.

From the FOAM GWA, significant association with plant height was observed at 46 unique hormonal pathway genes (Table 4.2), a significant enrichment of 64% compared to expected overlap by chance alone (permuted p-value ¡ 0.05). Of the 46 genes, 37 were associated to a single hormonal pathways. The highest representation was for auxin and ethylene, with a total of 8 genes each, followed by 6 genes related to abscisic acid, 5 to brassinosteroids, 4 for salicylic acid, and 3 for each cytokinin and gibberellin. Some of the corresponding best hit in *Arabidopsis* of the genes with significant association in the FOAM population are known for mutations causing height-related phenotypes. These included the *Arabidopsis* gene DWARF 3, a cytochrome P450 and is involved in brassinosteroids biosynthesis[25]; GA REQUIRING 1, whose mutation in *Arabidopsis* also leads dwarfing and which catalyzes the conversion of geranylgeranyl pyrophosphate (GGPP) to copalyl pyrophosphate (CPP) of gibberellin biosynthesis[26]. In addition, the list includes: CRY1/ CRYPTOCHROME 1, also known due to a strong allele as HY4/ ELONGATED HYPOCOTYL 4, which encodes a flavin-type blue-light photoreceptor [27]; CPD/ CONSTITUTIVE PHOTOMORPHOGENIC DWARF and CYP90D1/ CYTOCHROME P450, which encode cytochromes involved in brassinosteroid biosynthesis[25,28]; PID/PINOID, which encodes a kinase that may act as a positive regulator of cellular auxin efflux and as a negative regulator of auxin signaling[29]; DFL1/ DWARF IN LIGHT 1 which in *Arabidopsis* encodes an IAA-amido synthase involved in auxin homeostasis[30]; BIN2/ BRASSINOSTEROID-INSENSITIVE 2, also known as DWARF 12, which encodes a member of an *Arabidopsis* subfam-

ily of glycogen synthase kinase 3 and functions in the brassinosteroid signaling pathway[31]. The gene GRMZM2G092604, known in maize and *Arabidopsis* as bri1-like receptor kinase1, encodes in maize a brassinosteroid insensitive1-like receptor kinase, with recent experiments showing zmbri1-RNAi displays dwarfing phenotype [32].

In addition to height related phenotypes, we observed significant association at genes related with osmotic stress response, including the maize homologs of the *Arabidopsis* genes CIPK3/ CBL-INTERACTING PROTEIN KINASE 3, CBF2, and HAI2/ HIGHLY ABA-INDUCED PP2C GENE 2. In *Arabidopsis*, CIPK3 encodes a serine-threonine protein kinase that regulates ABA during germination, as well as having increased expression in cold, drought, high salt, and wounding conditions[33]. Similarly, the gene CBF2 encodes a member of the DREB subfamily A-1 of ERF/AP2 transcription factor family, and is involved in response to low temperature, abscisic acid, and circadian rhythm[34]. Finally, HAI2 encodes a member of the group A protein phosphatase 2C (PP2C) family that is responsible for negatively regulating seed dormancy[35], annotated in maize as ZmPP2C-A1. A study of natural variation in maize at the clade A PP2C phosphatases revealed a closely related protein, ZmPP2C-A10 to be associated with drought tolerance[36].

We looked at the most significant genes genome wide (Table 4.3), all with -log10 p-value greater than 24. This list includes the gene gt1 or grassy tillers1. This gene controls lateral branching, plant architecture and apical dominance in maize operating downstream of a key domestication gene, tb1, and gt1 was an important target of selection during domestication [37]. When looking further at the most significant marker at gt1, a pattern was observed where the

minor allele at low elevation is present at much higher frequency at high elevation, becoming the major allele (Figure 4.2). Interestingly, the allele with highest frequency at high elevation l also displays a negative effect on height from the association mode. In addition, to gt1 there was significance at several transcription factors including myb73 - MYB-transcription factor 73, a member of the very diverse family of myb proteins, which have been implicated in the regulation of a very wide variety of processes in plants including developmental processes[38]; mads69 - MADS-transcription factor 69, a member of the MADS-box family; and the gene gbptf1 GeBP-transcription factor 1. Members of the GeBP family have important role in hormonal pathway response[39], and the MADS-box gene family of transcription factors is present in many eukaryotes and in plants it includes the gene APETALA1 (AP1), which is involved in floral development.

In addition to overlap with genes, we compared the minor allele frequency distribution between all the segregating markers and those significantly associated with plant height. In a manner parallel yet more pronounced to flowering time in the same panel, we observed a minor allele frequency distribution highly enriched for low frequency polymorphisms, just above the minor allele frequency threshold for genome wide association (Figure 4.3). This is in sharp contrast to adaptive alleles associated with altitude or latitude adaptation, which are enriched at high minor allele frequencies.

Genomic selection

Genomic prediction are important models used to make phenotypic prediction and selections using markers. For flowering time in the same panel, a low

marker density corresponding to the top association markers (less than 1% of the markers) was observed to display as good predictive abilities as a much larger set of SNPs (30,000 of the total 500,000 segregating markers). For plant height, cross-validated predictive ability was estimated for each trials at increasing marker densities (Methods). Unlike flowering time, we observed that even with all half a million segregating markers, we obtained limited prediction accuracies across trials, with values ranging from 0.05 to 0.5, and a mean across trials of 0.25 (Figure 4.4). This is almost half of the prediction accuracy achieved for flowering time with less than 900 markers. The lowest prediction accuracies were obtained for trials subject to abiotic stress conditions.

4.4 Discussion

Maize landraces display a wide distribution of height values, a highly important evolutionary and agronomic trait. In terms of genetic architecture, in maize inbred lines plant height is a complex trait controlled by many genes of small effect[18]. Furthermore, the presence across many species[2,8], including maize[5,6,10,11,32], of dwarf mutants has lead to the elucidation of multiple hormonal pathways implicated in the control of plant height. Although landraces generally capture most of the genetic diversity of crop species, the genetic basis for height variation in maize farmers varieties remains largely uncharacterized. Here, we performed genome wide association and genomic prediction for plant height in a diverse panel of 4,100 landraces from Latin America evaluated across 30 trials.

In order to examine the association mapping results, we assembled multiple

lists of candidate genomic regions according to their biological evidence for potential role in height variation. First, we compared the significant SNPs to the regions of the genome which display low recombination in the same panel, as well as recent introgressions[20]. Because height displays high levels of heterosis, a significant overlap was expected between low recombination regions and height variation, however low overlap was observed compared to the overlap observed for flowering time. The use of a hybrid parent for the generation of the FOAM population and subsequent measurement across segregating progeny may have affected the ability to detect dominance effects. Nevertheless, the centromeres of chromosomes 3 and 5, as well as the putative chromosome 3 introgression were significantly associated with plant height showing a potential role for complementation of deleterious mutations at those three loci.

In the genic and recombining portion of the genome, we observed in total 1056 genes to be significantly associated to plant height variation across all trials, with the most significant ontology enrichment corresponding to anatomical structure architecture. For the genic hits, we looked first at the overlap between the association mapping results in inbred lines[18] and landraces. In total 95 and 41 genes overlapped between the FOAM results and the NAM and NCRPIS panels respectively. In particular, among the significant loci were ZCN8[24] and VGT1[40], two QTL implicated in flowering time variation. Furthermore, we compared the FOAM height and flowering time association results, and a three-fold enrichment was observed. This is consistent with the observed correlation between both flowering time and plant height across most measured trials. Although results from the NAM mapping show that genetically flowering time and plant height can be seen as independent traits[18], in diverse germplasm, like the NCRPIS inbred panel, and the landraces from the FOAM

panel, both traits may have been co-selected, inducing the genetic correlation.

An additional list of candidate genes assembled consisted of the best hits in maize of genes involved in 8 hormonal biosynthesis and signal transduction pathways in *Arabidopsis*[22]. We observed a 64% significant enrichment over the null expectation when comparing the significant genes from the FOAM GWA against the candidate gene list of hormone related genes. Some of the significant genes include classical maize and *Arabidopsis* genes displaying mutations leading to dwarfing phenotypes. In addition, several of the significant genes are implicated in *Arabidopsis* in stress response, including drought and freeze tolerance. Further research could help characterize standing genetic variation at those loci in maize landraces, as well as the mechanisms underlying their potential role in stress response.

Among the most significant genes we observed association at several transcription factors, suggesting a role regulation of gene expression underlying the genetic control of plant height. Furthermore, one of the most significant genes corresponded to the domestication locus grassy tillers1 (gt1). The gt1 gene affects apical dominance and plant architecture, acting as part of the shade avoidance response pathway as well as downstream of the domestication locus teosinte branched1. For gt1, we observed a unique minor allele frequency distribution, with the height reducing allele being most frequent at landrace native to high elevation locations; landraces from the Mexican highlands also present undesirable high levels of tillering[41], suggesting a potential role for natural standing variation at the gt1 in height and tillering variation in highland landraces, as well as a potential selection scheme to purge high tillering alleles from highland germplasm.

We compared the minor allele frequency distributions of all markers, as well as those significantly associated with plant height and flowering time. Most of the density for the height associated polymorphisms lies at low minor allele frequency, just above the minor allele frequency threshold for our genome wide association. This suggests that many rare alleles are likely affecting this trait, yet remain uncharacterized by association mapping approaches due to the loss in statistical power linked to the lack of allelic replication, an observation similar to association results in human stature[42]. Furthermore, given the high levels of heterosis observed for plant height this suggests that natural standing variation for plant height may be enriched in deleterious mutations, kept at low frequency under background selection. A significant reduction of effective population size, like the one experienced in the transition from landraces to breeding material, may lead the fixation of numerous divergent deleterious mutation affecting plant height across germplasm pools, which has been suggested would explain the observed heterotic effect in hybrids through complementation.

In terms of predictive ability, genomic selection displayed very modest prediction accuracies across trials, with increased values observed at higher marker densities. Trials with abiotic stress displayed the lowest prediction accuracies, however even using the entire set of segregating markers for trials under standard agronomic practices, prediction accuracies remained lower than those observed for flowering time. This suggests that for highly polygenic traits, greater marker density may be required in order increase prediction accuracies. In addition, the experimental design using a hybrid parent leads no dominance effects, and all non-additive effects remain untested with our current methodology. Alternative experimental designs along with statistical approaches incorporating

dominance effects could help unveil the role of dominance in plant height, as well as improve prediction accuracies.

Because of their long recombination history and high diversity, landraces offer an excellent system to find candidate loci associated with quantitative traits at high resolution. The significant enrichment with genes implicated in regulation and anatomical structure lay a foundation to explore the mechanisms underlying the developmental control of plant height. The use of whole genome sequencing, or targeted sequencing of loci at key pathways affecting plant height or abiotic stress response could allow the application of genome editing to test and manipulate highly polygenic traits like drought response and plant height, allowing the fast the movement of alleles across populations.

4.5 Methods

Phenotypic evaluation

We used individuals from the maize landrace FOAM population to study the genetic basis of height variation[20]. Briefly, this population represents the F1 progeny from 4,500 landrace individuals crossed to a limited number of hybrid parents. The design for crossing and evaluation of maize landraces is nested within 3 previously defined[43] adaptation classes: lowland tropical (¡1200 meters above sea level and ¡30 N or 40 S). subtropical (between 1200 and 1900 m.a.s.l. and ¡30 N or 40 S), and highland (above 1900 m.a.s.l. and ¡30 N or 40 S). This nested design allows accurate evaluation of landraces, avoiding confounding with lack of adaptation during field experiments.

In total 30 trials over 2 years in 13 locations across Mexico were grownusing an extended row column design[44], with accessions tested only in locations of matching adaptation. Each trial contained between 292 and 2082 accessions, with an average of 887. There were between 10-15 plants per row per accession. For each trial, plant height was measured on 5 plants per row from the base of the plant (where the stalk comes out of the soil) to the base of the tassel. Standard agronomic conditions were used for 24 of the trials, and 6 trials were conducted under abiotic stress conditions (Table 4.8). The stress trials were exact replicas of other trials conducted under standard agronomic conditions in the same location and year. The stress trials were 4 trials for which low Nitrogen fertilizer was used, and 2 trials for which low irrigations was used.

Phenotypic data was analysed using a mixed model to estimate the genetic contribution of the landrace donors for a total of 26,634 breeding values. The model used was the same as reported for previous analyses on the same panel[20] and includes fixed effects for checks, tester, and hybrid and a random effect of accession in a complete nested model. Including in the model the random effect of row and column and using an autoregressive model of order 1 in row and columns controlled experimental noise product of field variation. All the random effects were considered independent one from each other.

Candidate genes

We obtained the list of high-LD features of the maize genome present in the maize FOAM landrace panel[20]. This regions represent the centromeres and pericentromeric regions, as well as features like chromosomal inversions. Because of extended LD, these regions are not suitable for candidate gene analy-

ses. Using Ensembl BioMarts[45] we assembled a list of 1035 homologous genes in the maize genome.

Genome wide association

We performed genome wide association with plant height using the BEAGLE4[46] imputed Genotyping-by-Sequencing (GBS) markers generated for the FOAM landrace parents. Markers were filtered for minor allele frequency greater than 1%. We used the same model reported for studying flowering time in the same panel[20], which fits the GWA model in two steps. First, a model is fit accounting for the main effects of trial, hybrid, population structure through 10 Multidimensional Scaling weights, and a random effect of relatedness through a kinship matrix. The second model fits the SNP effect on the residuals from the first model, with SNP being nested within trial to test for significant genotypic effect on any trial. The resulting p-values are filtered with a threshold of significance of -log10 greater than 9.

Genomic selection

In order to explore the potential for selection using genomewide markers, we performed genomic prediction for plant height using by running 20 iteration, each estimating a 5-fold cross validated prediction accuracy using the software GAPIT[47]. We estimated the prediction accuracy for each trial with models accounting for population structure in the form of 10 multidimensional scaling weights, and using kinship matrices estimated using either 30000 or 280000 random markers, or all 500000 segregating markers.

4.6 References

1. Smith H, Whitelam GC. The shade avoidance syndrome: multiple responses mediated by multiple phytochromes. Plant Cell Environ. Blackwell Publishing Ltd; 1997;20: 840844.

2. Peng J, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, et al. Green revolution genes encode mutant gibberellin response modulators. Nature. 1999;400: 256261.

3. Sasaki A, Ashikari M, Ueguchi-Tanaka M, Itoh H, Nishimura A, Swapan D, et al. Green revolution: a mutant gibberellin-synthesis gene in rice. Nature. 2002;416: 701702.

4. Hedden P. The genes of the Green Revolution. Trends Genet. 2003;19: 59.

5. Phinney BO. GROWTH RESPONSE OF SINGLE-GENE DWARF MUTANTS IN MAIZE TO GIBBERELLIC ACID. Proc Natl Acad Sci U S A. 1956;42: 185189.

6. Multani DS, Briggs SP, Chamberlin MA, Blakeslee JJ, Murphy AS, Johal GS. Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. Science. 2003;302: 8184.

7. Noguchi T, Fujioka S, Choe S, Takatsuto S, Yoshida S, Yuan H, et al. Brassinosteroid-insensitive dwarf mutants of Arabidopsis accumulate brassinosteroids. Plant Physiol. 1999;121: 743752.

8. Ishikawa S, Maekawa M, Arite T, Onishi K, Takamure I, Kyozuka J. Suppression of tiller bud activity in tillering dwarf mutants of rice. Plant Cell Physiol. 2005;46: 7986.

9. Bensen RJ, Johal GS, Crane AVC, Tossberg JT, Schnable PS. Cloning and Characterization of the Maize Anl Gene. Plant Cell. 1995;7: 7584.

10. Fujioka S, Yamane H, Spray CR, Gaskin P, Macmillan J, Phinney BO, et al.

Qualitative and Quantitative Analyses of Gibberellins in Vegetative Shoots of Normal, dwarf-1, dwarf-2, dwarf-3, and dwarf-5 Seedlings of Zea mays L. Plant Physiol. 1988;88: 13671372.

11. Winkler RG, Helentjaris T. The Maize Dwarf3 Gene Encodes a Cytochrome P450-Mediated Early Step in Gibberellin Biosynthesis. Plant Cell. 1995;7: 13071317.

12. Lawit SJ, Wych HM, Xu D, Kundu S, Tomes DT. Maize DELLA proteins dwarf plant8 and dwarf plant9 as modulators of plant development. Plant Cell Physiol. 2010;51: 18541868.

13. Ueguchi-Tanaka M, Ashikari M, Nakajima M, Itoh H, Katoh E, Kobayashi M, et al. GIBBERELLIN INSENSITIVE DWARF1 encodes a soluble receptor for gibberellin. Nature. Nature Publishing Group; 2005;437: 693698.

14. Zhao Y. Auxin biosynthesis and its role in plant development. Annu Rev Plant Biol. 2010;61: 4964.

15. Bao F, Shen J, Brady SR, Muday GK, Asami T, Yang Z. Brassinosteroids interact with auxin to promote lateral root development in Arabidopsis. Plant Physiol. 2004;134: 16241631.

16. Hartwig T, Chuck GS, Fujioka S, Klempien A, Weizbauer R, Potluri DPV, et al. Brassinosteroid control of sex determination in maize. Proc Natl Acad Sci U S A. 2011;108: 1981419819.

17. Salas Fernandez MG, Becraft PW, Yin Y, Lbberstedt T. From dwarves to giants? Plant height manipulation for biomass yield. Trends Plant Sci. 2009;14: 454461.

18. Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, et al. The genetic architecture of maize height. Genetics. 2014;196: 13371356.

19. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining

the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46: 11731186.

20. Alberto Romero Navarro J, Wilcox M, Burgueo J, Romay C, Swarts K, Trachsel S, et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. Nat Genet. Nature Research; 2017; doi:10.1038/ng.3784

21. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013;14: R55. 22. Peng Z-Y, Zhou X, Li L, Yu X, Li H, Jiang Z, et al. Arabidopsis Hormone Database: a comprehensive genetic and phenotypic information database for plant hormone research in Arabidopsis. Nucleic Acids Res. 2009;37: D97582.

23. Hufford MB, Lubinksy P, Pyhjrvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013;9: e1003477.

24. Meng X, Muszynski MG, Danilevskaya ON. The FT-like ZCN8 Gene Functions as a Floral Activator and Is Involved in Photoperiod Sensitivity in Maize. Plant Cell. 2011;23: 942960.

25. Szekeres M, Nmeth K, Koncz-Klmn Z, Mathur J, Kauschmann A, Altmann T, et al. Brassinosteroids rescue the deficiency of CYP90, a cytochrome P450, controlling cell elongation and de-etiolation in Arabidopsis. Cell. 1996;85: 171182.

26. Sun TP, Kamiya Y. The Arabidopsis GA1 locus encodes the cyclase ent-kaurene synthetase A of gibberellin biosynthesis. Plant Cell. 1994;6: 15091518.

27. Ahmad M, Cashmore AR. HY4 gene of A. thaliana encodes a protein with characteristics of a blue-light photoreceptor. Nature. 1993;366: 162166.

28. Kim G-T, Fujioka S, Kozuka T, Tax FE, Takatsuto S, Yoshida S, et al. CYP90C1 and CYP90D1 are involved in different steps in the brassinosteroid biosynthe-

sis pathway in Arabidopsis thaliana. Plant J. Wiley Online Library; 2005;41: 710721.

29. Christensen SK, Dagenais N, Chory J, Weigel D. Regulation of auxin response by the protein kinase PINOID. Cell. 2000;100: 469478.

30. Staswick PE, Serban B, Rowe M, Tiryaki I, Maldonado MT, Maldonado MC, et al. Characterization of an Arabidopsis enzyme family that conjugates amino acids to indole-3-acetic acid. Plant Cell. 2005;17: 616627.

31. Belkhadir Y, Chory J. Brassinosteroid signaling: a paradigm for steroid hormone signaling from the cell surface. Science. 2006;314: 14101411.

32. Kir G, Ye H, Nelissen H, Neelakandan AK, Kusnandar AS, Luo A, et al. RNA Interference Knockdown of BRASSINOSTEROID INSENSITIVE1 in Maize Reveals Novel Functions for Brassinosteroid Signaling in Controlling Plant Architecture. Plant Physiol. 2015;169: 826839.

33. Kim K-N, Cheong YH, Grant JJ, Pandey GK, Luan S. CIPK3, a calcium sensor-associated protein kinase that regulates abscisic acid and cold signal transduction in Arabidopsis. Plant Cell. 2003;15: 411423.

34. Medina J, Catal R, Salinas J. The CBFs: three arabidopsis transcription factors to cold acclimate. Plant Sci. 2011;180: 311.

35. Nakashima K, Yamaguchi-Shinozaki K. ABA signaling in stress-response and seed development. Plant Cell Rep. 2013;32: 959970.

36. Xiang Y, Sun X, Gao S, Qin F, Dai M. Deletion of an Endoplasmic Reticulum Stress Response Element in a ZmPP2C-A Gene Facilitates Drought Tolerance of Maize Seedlings. Mol Plant. 2017;10: 456469.

37. Whipple CJ, Kebrom TH, Weber AL, Yang F, Hall D, Meeley R, et al. grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. Proc Natl Acad Sci U S A. 2011;108: E50612.

38. Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L. MYB transcription factors in Arabidopsis. Trends Plant Sci. 2010;15: 573581.

39. Chevalier F, Perazza D, Laporte F, Le Hnanff G, Hornitschek P, Bonneville J-M, et al. GeBP and GeBP-like proteins are noncanonical leucine-zipper transcription factors that regulate cytokinin response in Arabidopsis. Plant Physiol. 2008;146: 11421154.

40. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci U S A. 2007;104: 1137611381.

41. Perales R. H, Brush SB, Qualset CO. Landraces of Maize in Central Mexico: An Altitudinal Transect. Econ Bot. The New York Botanical Garden; 2003;57: 720.

42. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. 2017;542: 186190.

43. Salhuana W, Jones Q, Sevilla R. The Latin American Maize Project: Model for rescue and use of irreplaceable germplasm. Diversity . 1991;

44. Crossa J, Federer WT. I.4 Screening Experimental Designs for Quantitative Trait Loci, Association Mapping, Genotype-by Environment Interaction, and Other Investigations. Front Physiol. Frontiers; 2012;3. doi:10.3389/fphys.2012.00156

45. Kinsella RJ, Khri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database . 2011;2011: bar030.

46. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics. 2013;194: 459471.

47. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28: 23972399.

4.7 Figures

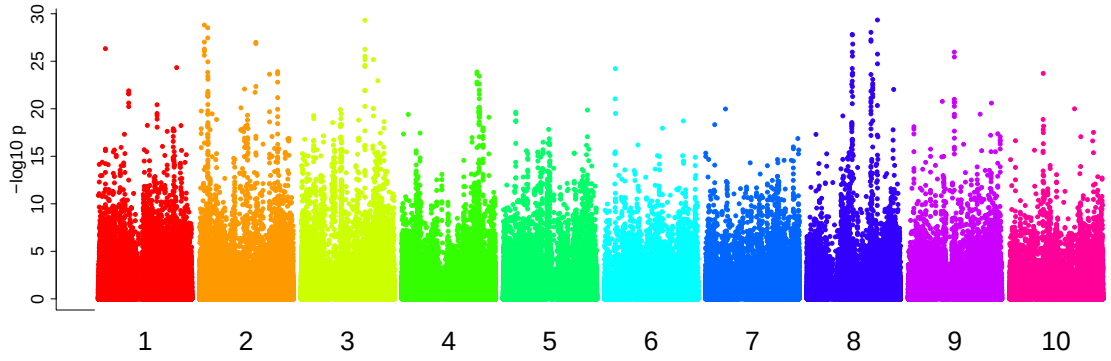


Figure 4.1: Manhattan plot of genome wide association with plant height

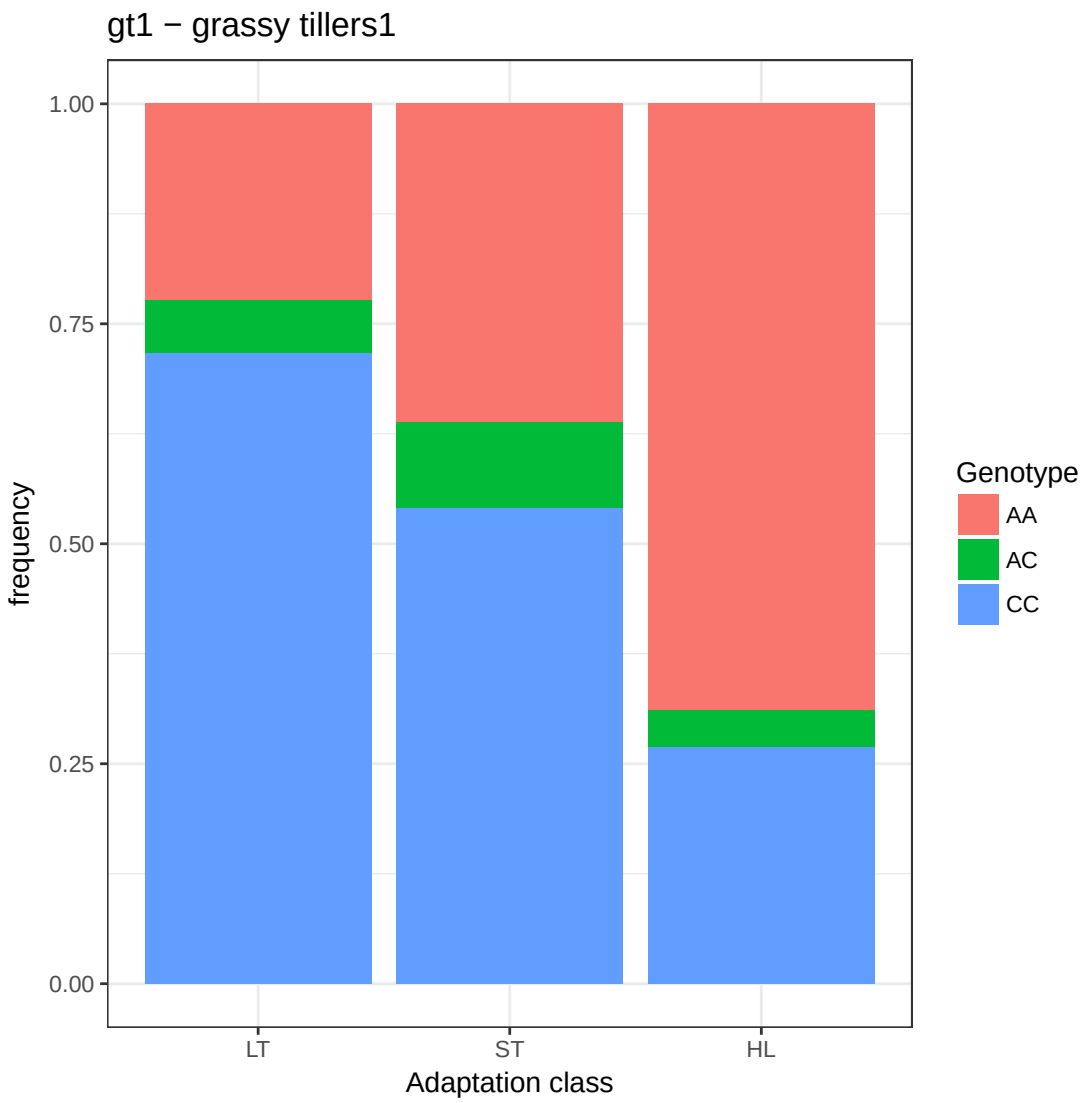


Figure 4.2: Genotype frequencies of the most significant marker at the grassy tillers 1 gene by adaptation zone

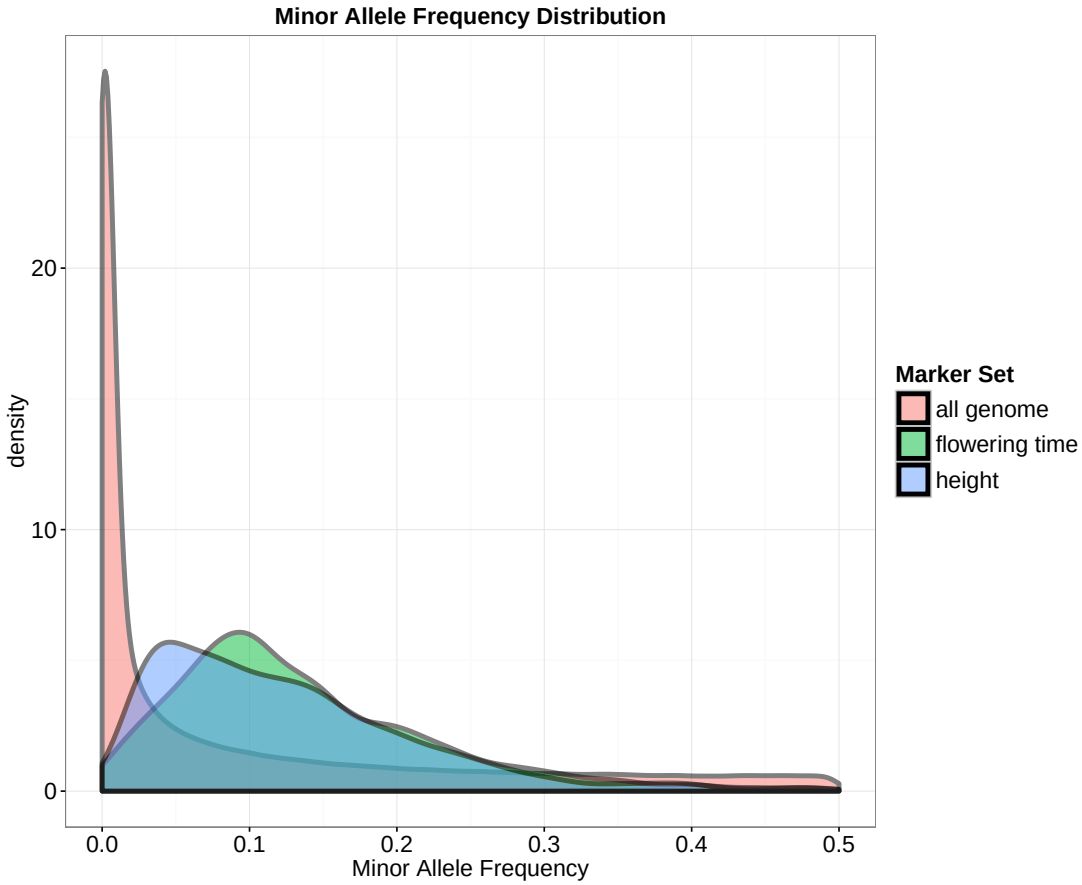


Figure 4.3: Minor Allele Frequency of height associated SNPs

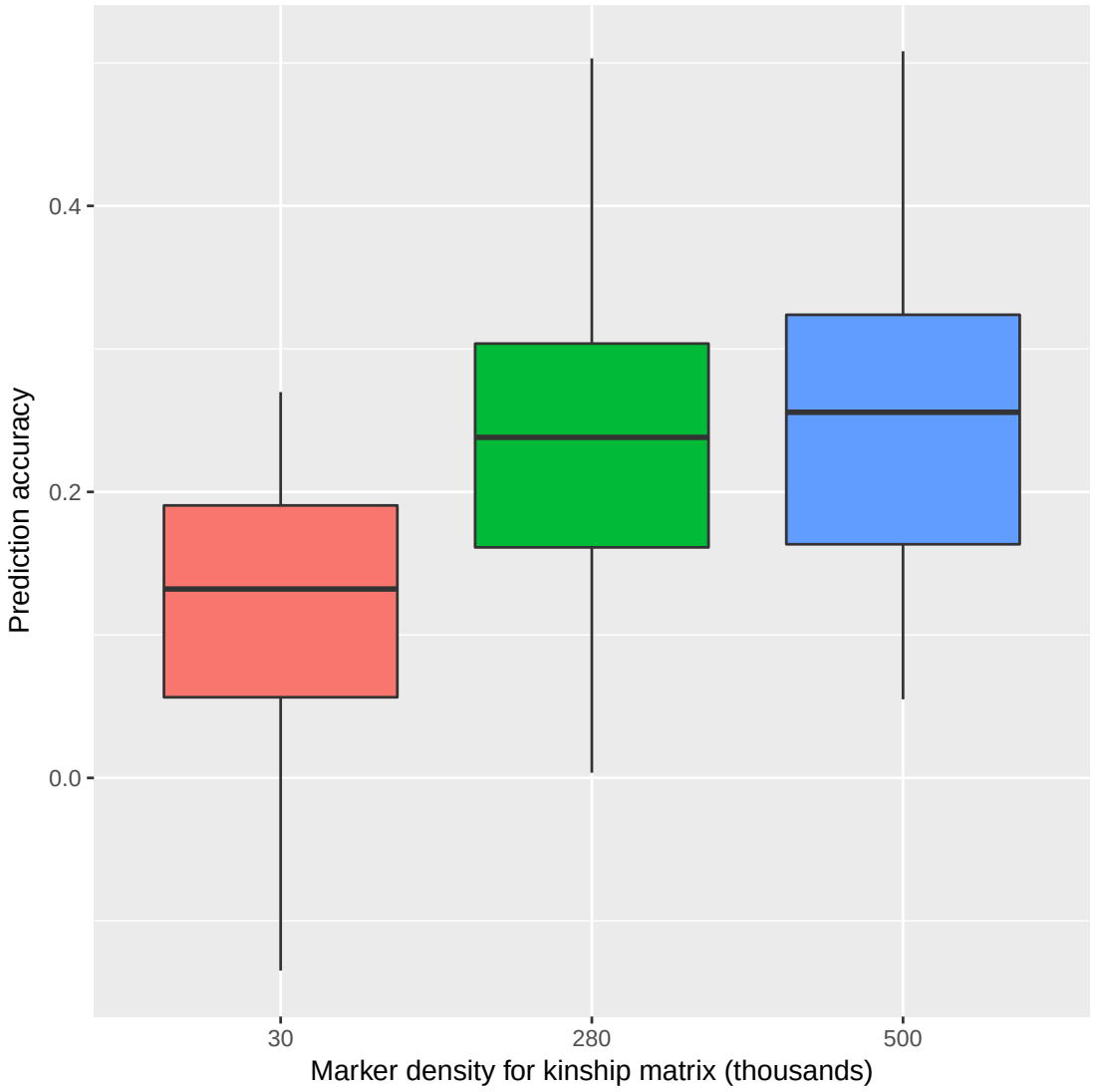


Figure 4.4: Prediction accuracy for plant height across marker densities

4.8 Tables

Trial name	Year	Location	State	Accessions	Heritability	Agronomics
m2011BAFNB	2011	Agua Fria	Puebla	1871	44.6	Low Nitrogen
m2011BAFNN	2011	Agua Fria	Puebla	1870	61.1	Standard
m2011BCH	2011	G Victoria	Chiapas	510	33.9	Standard
m2011BCL	2011	Celaya	Guanajuato	829	42.9	Standard
m2011BMO	2011	Tarimbaro	Michoacan	515	56.7	Standard
m2011BNY	2011	San Pedro	Nayarit	592	61.4	Standard
m2011BTLNB	2011	Tlaltizapan	Morelos	1217	19.6	Low Nitrogen
m2011BTLNN	2011	Tlaltizapan	Morelos	1217	67.9	Standard
m2011BTORN	2011	Torreon	Coahuila	1517	37.3	Standard
m2012AAFCA	2012	Agua Fria	Puebla	2078	71.1	Standard
m2012AAFTU	2012	Agua Fria	Puebla	2082	61.3	Standard
m2012AIGRN	2012	Iguala	Guerrero	873	53.9	Standard
m2012AOBES	2012	Obregon	Sonora	484	18.7	Standard
m2012AOBRN	2012	Obregon	Sonora	484	55.6	Standard
m2012BAFNB	2012	Agua Fria	Puebla	726	44.8	Low Nitrogen
m2012BAFNN	2012	Agua Fria	Puebla	731	46.1	Standard
m2012BAL	2012	Almoloya	Mexico	587	49.4	Standard
m2012BBACA	2012	El Batan	Mexico	815	73.7	Standard
m2012BBAEF	2012	El Batan	Mexico	959	47.2	Standard
m2012BBANB	2012	El Batan	Mexico	816	54.3	Low Nitrogen
m2012BBANN	2012	El Batan	Mexico	816	59.0	Standard
m2012BCH	2012	G Victoria	Chiapas	680	66.1	Standard
m2012BCL	2012	Celaya	Guanajuato	814	83.5	Standard
m2012BEB	2012	Cortazar	Guanajuato	669	80.5	Standard
m2012BMO	2012	Numaran	Michoacan	292	66.5	Standard
m2012BNY	2012	San Pedro	Nayarit	813	22.7	Standard
m2012BOBRN	2012	Obregon	Sonora	543	77.6	Standard
m2012BOBRR	2012	Obregon	Sonora	543	68.6	Low Irrigation
m2012BTORN	2012	Torreon	Coahuila	345	58.9	Standard
m2012BTORR	2012	Torreon	Coahuila	346	31.6	Low Irrigation

Table 4.1: Plant height trial information

Chr	Position	Maize gene ID / name	Hormone	*Arabidopsis* best hit	*Arabidopsis* annotation
1	23241091-23244476	GRMZM2G005624/ grassy tillers 1	Brassinosteroid	HB-2/ HOMEOBOX-2	Homeodomain leucine zipper class I (HD-Zip I) protein
1	283088437-283091230	GRMZM2G065214	Abscisic acid	PERK4/ PROLINE-RICH EXTENSIN-LIKE RECEPTOR KINASE 4	Proline-rich extensin-like receptor kinase 4. Functions at an early stage of ABA signalling inhibiting primary root cell elongation by perturbing Ca2+ homeostasis
1	241218603-241226854	GRMZM2G081554/ anther ear1	Gibberellin	GA1/ GA REQUIRING 1	Catalyzes the conversion of geranylgeranyl pyrophosphate (GGPP) to copalyl pyrophosphate (CPP) of gibberellin biosynthesis
1	189351713-189357321	GRMZM2G124557	Brassinosteroid	AtRGTB2/ RAB GERANYLGERANYL TRANSFERASE BETA SUBUNIT 2	RAB geranylgeranyl transferase beta subunit 2
1	287719533-287722437	GRMZM2G133023	Gibberellin	AT3G22850.1	Aluminum induced protein with YGL and LRDR motifs
1	38124980-38130704	GRMZM2G161337	Auxin / Jasmonic acid	ASA2/ ANTHRANILATE SYNTHASE 2	Encodes a functional anthranilate synthase protein. Involved in aromatic amino acid biosynthesis
1	34621918-34624942	GRMZM2G171736	Auxin / Salicylic acid	CRY1/ CRYPTOCHROME 1	Encodes a flavin-type blue-light photoreceptor with ATP binding and autophosphorylation activity. The photoreceptor may be involved in electron transport. Mutant phenotype displays a blue light-dependent inhibition of hypocotyl elongation
1	173846756-173848499	GRMZM2G171822/ barren inflorescence2	Auxin	PID/ PINOID	Encodes a protein serine/threonine kinase that may act as a positive regulator of cellular auxin efflux, as a a binary switch for PIN polarity, and as a negative regulator of auxin signaling. Recessive mutants exhibit similar phenotypes as pin-formed mutants in flowers and inflorescence but distinct phenotypes in cotyledons and leaves

1	52150727-52155845	GRMZM2G174896/ calcineurin B-like-interacting protein kinase3	Abscisic acid	CIPK3/ CBL-INTERACTING PROTEIN KINASE 3	Encodes a serine-threonine protein kinase whose expression increases in response to abscisic acid, cold, drought, high salt, and wounding conditions
1	154952735-154957201	GRMZM2G403620/ rough sheath2	Gibberellin	ATPHAN/ ARABIDOPSIS PHANTASTICA-LIKE 1	Encodes a MYB-domain protein involved in specification of the leaf proximodistal axis. Mutation results in lobed and dissected leaves with a characteristic asymmetry
2	2266715-2274671	GRMZM2G078274/ ARF-transcription factor 3	Auxin	ARF6/ AUXIN RESPONSE FACTOR 6	Encodes a member of the auxin response factor family. Mediates auxin response via expression of auxin regulated genes. Acts redundantly with ARF8 to control stamen elongation and flower maturatio.
2	9445775-9447650	GRMZM2G080054/ bHLH-transcription factor 148	Abscisic acid / Gibberellin	PIL5/ PHYTOCHROME INTERACTING FACTOR 3-LIKE 5	Encodes a Myc-related bHLH transcription factor that has transcriptional activation activity in the dark. It is a key negative regulator of phytochrome-mediated seed germination and acts by inhibiting chlorophyll biosynthesis, light-mediated suppression of hypocotyl elongation and far-red light-mediated suppression of seed germination, and promoting negative gravitropism in hypocotyls
2	131348486-131350096	GRMZM2G124715/ MYB-transcription factor 64	Abscisic acid	MYB4/ MYB DOMAIN PROTEIN 4	Encodes a R2R3 MYB protein which is involved in the response to UV-B
2	146049603-146052033	GRMZM2G162737	Brassinosteroid	CPD/ CONSTITUTIVE PHOTOMORPHOGENIC DWARF	Encodes a member of the CP90A family, a cytochrome P450 monooxygenase which converts 6-deoxocathasterone to 6-deoxoteasterone in the late C6 oxidation pathway and cathasterone to teasterone in the early C6 oxidation pathway of brassinolide biosynthesis. Mutants display de-etiolation and derepression of light-induced genes in the dark, dwarfism, male sterility and activation of stress-regulated genes in the light

2	12251240-12258971	GRMZM2G374203	Brassinosteroid / cytokinin	BRX/ BREVIS RADIX	Belongs to five-member BRX gene family. BRX encodes a key regulator of cell proliferation and elongation in the root, which has been implicated in the brassinosteroid (BR) pathway as well as regulation of auxin-responsive gene expression
3	196568222-196574187	GRMZM2G056120/ ARF-transcription factor 11	Auxin	ARF3/ AUXIN RESPONSE TRANSCRIPTION FACTOR 3	Mutations have pleiotropic effects on Arabidopsis flower development, causing increases in perianth organ number, decreases in stamen number and anther formation, and apical-basal patterning defects in the gynoecium. The gene encodes a protein with homology to DNA binding proteins which bind to auxin response elements
3	180041545-180043505	GRMZM2G059453/ protein phosphatase homolog3	Abscisic acid	HAI2/ HIGHLY ABA-INDUCED PP2C GENE 2	Encodes a member of the group A protein phosphatase 2C (PP2C) family that is responsible for negatively regulating seed dormancy
3	50093119-50099618	GRMZM2G119894	Auxin	PGP21/ P-GLYCOPROTEIN 21	Encodes a facultative transporter controlling auxin concentrations in plant cells
3	178602863-178606881	GRMZM2G125411	Abscisic acid / Jasmonic acid	COI1/ CORONATINE INSENSITIVE 1	Encodes a protein containing Leu-rich repeats and a degenerate F-box motif. Associates with AtCUL1, AtRbx1, and the Skp1-like proteins ASK1 and ASK2 to assemble SCF COI1 ubiquitin-ligase complexes in planta. A single amino acid substitution in the F-box motif of COI1 abolishes the formation of the SCF(COI1) complexes and results in loss of the JA response
3	5580132-5586653	GRMZM2G143235	Brassinosteroid	CYP90D1/ CYTOCHROME P450, FAMILY 90, SUBFAMILY D, POLYPEPTIDE 1	Encodes a cytochrome P-450 gene that is involved in brassinosteroid biosynthesis
3	188669571-188670618	GRMZM2G310368/ AP2-EREBP-transcription factor 147	Gibberellin	ERF9/ ERF DOMAIN PROTEIN 9	Encodes a member of the ERF (ethylene response factor) subfamily B-1 of ERF/AP2 transcription factor family (ATERF-9). The protein contains one AP2 domain

3	9303333-9314353	GRMZM5G844173/ gigantea2	Gibberellin	GI/ GIGANTEA	Together with CONSTANTS (CO) and FLOWERING LOCUS T (FT), GIGANTEA promotes flowering under long days in a circadian clock-controlled flowering pathway
4	198130795-198133409	GRMZM2G027972/ WRKY-transcription factor 87	Abscisic acid	WRKY2/ WRKY DNA-BINDING PROTEIN 2	Encodes WRKY transcription factor 2, a zinc-finger protein. In wrky2 mutants, egg cells polarize normally but zygotes fail to reestablish polar organelle positioning from a transient symmetric state, resulting in equal cell division and distorted embryo development
4	118695002-118700985	GRMZM2G137413/ ARF-transcription factor 14	Auxin	ARF1/ AUXIN RESPONSE FACTOR 1	Encodes a member of the auxin response factor family. ARFs bind to the cis element 5′-TGTCTC-3′ ARFs mediate changes in gene expression in response to auxin. ARF's form heterodimers with IAA/AUX genes. ARF1 enhances mutant phenotypes of ARF2 and may act with ARF2 to control aspects of maturation and senescence
4	230754420-230756943	GRMZM2G318689	Gibberellin	EIN4/ ETHYLENE INSENSITIVE 4	Ethylene receptor, subfamily 2. Has serine kinase activity
5	168251978-168264266	GRMZM2G059671	Abscisic acid / Auxin / Gibberellin	CTR1/ CONSTITUTIVE TRIPLE RESPONSE 1	Homologous to the RAF family of serine/threonine protein kinases. Negative regulator in the ethylene signal transduction pathway. Interacts with the putative ethylene receptors ETR1 and ERS
5	7772505-7775451	GRMZM2G077356/ Aux/IAA- transcription factor 21	Auxin	IAA16/ INDOLEACETIC ACID-INDUCED PROTEIN 16	Early auxin-induced (IAA16)
5	38752228-38809303	GRMZM2G158252/ histidine kinase3	Cytokinin	HK2/ HISTIDINE KINASE 2	Encodes histidine kinase AHK2

5	213732108-213736912	GRMZM2G479110/ ARR-B-transcription factor 8	Cytokinin	RR12/ RESPONSE REGULATOR 12	Encodes an Arabidopsis response regulator (ARR) protein that acts in concert with other type-B ARRs in the cytokinin signaling pathway. Also involved in cytokinin-dependent inhibition of hypocotyl elongation and cytokinin-dependent greening and shooting in tissue culture
6	74598823-74604163	GRMZM2G002100/ MAP kinase7	Abscisic acid / Gibberellin / Jasmonic acid	MPK6/ Salicylic acid MAP KINASE 6	Encodes a MAP kinase induced by pathogens, ethylene biosynthesis, oxidative stress and osmotic stress. Also involved in ovule development
6	37025125-37027411	GRMZM2G021909	Gibberellin	Epsin N-terminal homology (ENTH) domain-containing protein	Epsin N-terminal homology (ENTH) domain-containing protein / Clathrin assembly protein-like protein
7	7410114-7413460	GRMZM2G044469	Abscisic acid	AAO1/ ALDEHYDE OXIDASE 1	Encodes aldehyde oxidase AA01
7	53466716-53467324	GRMZM2G060517/ AP2-EREBP-transcription factor 189	Gibberellin	ERF96/ ETHYLENE RESPONSE FACTOR 96	Encodes a member of the ERF (ethylene response factor) subfamily B-3 of ERF/AP2 transcription factor family. The protein contains one AP2 domain. There are 18 members in this subfamily including ATERF-1, ATERF-2, AND ATERF-5. Expression of ERF96 is induced by pathogens, JA and ethylene and over expression leads to increased resistance to resistance to necrotrophic pathogens
7	141133700-141134759	GRMZM2G069126/ AP2-EREBP-transcription factor 23	Abscisic acid / Gibberellin / Jasmonic acid / Salicylic acid	CBF2/ C-REPEAT/ DRE BINDING FACTOR 2	Encodes a member of the DREB subfamily A-1 of ERF/AP2 transcription factor family (CBF2). The protein contains one AP2 domain. There are six members in this subfamily, including CBF1, CBF2, and CBF3. This gene is involved in response to low temperature, abscisic acid, and circadian rhythm. Overexpressing this gene leads to increased freeze tolerance and induces the expression level of 85 cold-induced genes and reduces the expression level of 8 cold-repressed genes, which constitute the CBF2 regulon

7	79095632-79100094	GRMZM2G092604/ bri1-like receptor kinase1	Brassinosteroid	BRL1/ BRI1 LIKE	Mutant has altered vascular cell differentiation
7	92391794-92393825	GRMZM2G139815/ WRKY-transcription factor 37	Salicylic acid	WRKY30/ WRKY DNA-BINDING PROTEIN 30	Member of WRKY Transcription Factor Group III
8	133181699-133186955	GRMZM2G005350	Salicylic acid	MKP1/ MITOGEN-ACTIVATED PROTEIN KINASE PHOSPHATASE 1	Encodes MAP kinase phosphatase 1 (MKP1). Loss of MKP1 results in hypersensitivity to acute UV-B stress, but without impairing UV-B acclimation
8	72382998-72384377	GRMZM2G063880/ WRKY-transcription factor 106	Salicylic acid	WRKY53	Member of WRKY Transcription Factor Group III
8	120135828-120138590	GRMZM2G080731/ bZIP-transcription factor 25	Salicylic acid	BZIP45	Basic leucine zipper transcription factor involved in the activation of SA-responsive genes
8	121797016-121799823	GRMZM2G366873	Auxin	DFL1/ DWARF IN LIGHT 1	Encodes an IAA-amido synthase that conjugates Ala, Asp, Phe, and Trp to auxin. Lines overexpressing this gene accumulate IAA-ASP and are hypersensitive to several auxins. Identified as a dominant mutation that displays shorter hypocotyls in light grown plants when compared to wild type siblings
8	151594820-151595908	GRMZM2G393014/ isopentenyl transferase3	Cytokinin	IPT1/ ISOPENTENYL-TRANSFERASE 1	Encodes a putative adenylate isopentenyltransferase. It catalyzes the formation of isopentenyladenosine 5′-monophosphate (iPMP) from AMP and dimethylallylpyrophosphate (DMAPP), but it has a lower Km for ADP and likely works using ADP or ATP in plants. It is involved in cytokinin biosynthesis
8	23422894-23427791	GRMZM2G472625	Brassinosteroid / Abscisic acid	BIN2/ BRASSINOSTEROID-INSENSITIVE 2	Encodes BIN2, a member of the ATSK (shaggy-like kinase) family. BIN2 functions in the cross-talk between auxin and brassinosteroid signaling pathways
9	149749178-149751820	GRMZM2G095786	Auxin	AFB4/ AUXIN SIGNALING F-BOX 4	RNI-like superfamily protein

10	75668946-75675094	GRMZM2G055125	Gibberellin	MOB1-like/ MOB1-LIKE	Encodes a gene product involved in both sporogenesis and gametogenesis and is required for the normal progression of megasporogenesis and microsporogenesis. Additional alleles were isolated in a screen for enhancers of PID and genetic analysis indicates a role for MOB1A in auxin mediated signaling
10	140190031-140191484	GRMZM2G097636/ MYB-transcription factor 5	Gibberellin	MYB63/ MYB DOMAIN PROTEIN 63	Member of the R2R3 factor gene family
10	148424594-148425826	GRMZM2G142802	Gibberellin	aluminum induced protein with YGL and LRDR motifs	aluminum induced protein with YGL and LRDR motifs

Table 4.2: Plant height genes orthologous to genes annotated in *Arabidopsis* hormonal pathways. Annotations from *Arabidopsis* were obtained from The Arabidopsis Information Resource (TAIR) database

Chr	start	Name	*Arabidopsis* gene	*Arabidopsis* Annotation
1	23241091	GRMZM2G005624 gt1 grassy tillers1	AT2G18550.1	(ATHB21, HB-2, HB21) homeobox protein 21
1	252320123	GRMZM2G121570 myb73	AT5G58850.1	(ATMYB119, MYB119) myb domain protein 119
2	12351628	GRMZM2G125728	AT5G11710.1	ENTH/VHS family protein
2	12437551	GRMZM2G313020	AT1G65810.1	P-loop containing nucleoside triphosphate hydrolases superfamily protein
2	21007540	AC212835.3_FG007	AT4G17100.1	CONTAINS InterPro DOMAIN/s: Endoribonuclease XendoU
2	21273857	GRMZM2G020701	AT2G19170.1	(SLP3) subtilisin-like serine protease 3
2	142624877	GRMZM2G144662	AT3G05330.1	(ATN, ATTAN) cyclin family
3	158979321	GRMZM2G171650	N/A	N/A
3	159169738	GRMZM2G082608 mads69 - MADS-transcription factor 69	AT2G03710.3	(AGL3, SEP4) K-box region and MADS-box transcription factor family protein
3	180041545	GRMZM2G059453	AT1G07430.1	(HAI2) highly ABA-induced PP2C gene 2
6	19962512	GRMZM2G392516 gbptf1 - GeBP-transcription factor 1	AT4G00610.1	DNA-binding storekeeper protein-related transcriptional regulator
8	85120975	GRMZM2G449496	AT3G46780.1	(PTAC16) plastid transcriptionally active 16
8	85144906	GRMZM2G132633	AT1G07480.1	Transcription factor IIA alpha/beta subunit
8	85306614	GRMZM2G044997	AT3G20420.1	(ATRTL2, RTL2) RNAse THREE-like protein 2
8	85398753	GRMZM2G060142	AT1G79890.1	RAD3-like DNA-binding helicase protein
8	86063170	GRMZM2G037012	AT3G06840.1	unknown protein
8	120149428	GRMZM2G036980	N/A	N/A
8	132327747	GRMZM2G140083 hb16 - Homeobox-transcription factor 16	AT2G28610.1	Homeodomain containing protein that regulates lateral axis-dependent development
8	132372125	GRMZM2G353250	N/A	N/A
9	77179962	GRMZM2G080907	AT5G44410.1	FAD-binding Berberine family protein
9	77309262	GRMZM2G110582 mads71 - MADS-transcription factor 71	AT5G60440.1	(AGL62) AGAMOUS-like 62

Table 4.3: Plant height top genes

4.9 Supplemental material

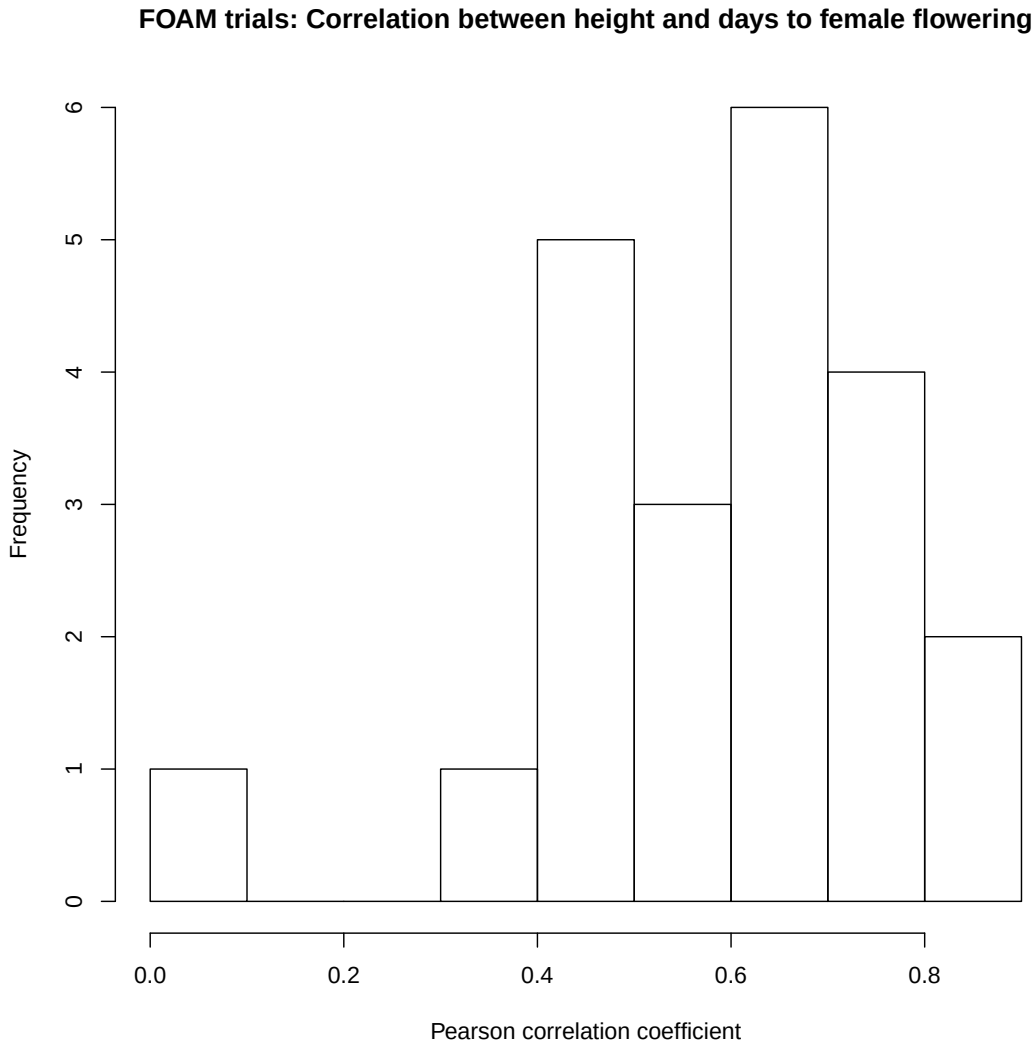


Figure 4.5: Distribution of Pearson correlation coefficients estimated between female flowering time and plant height across trials

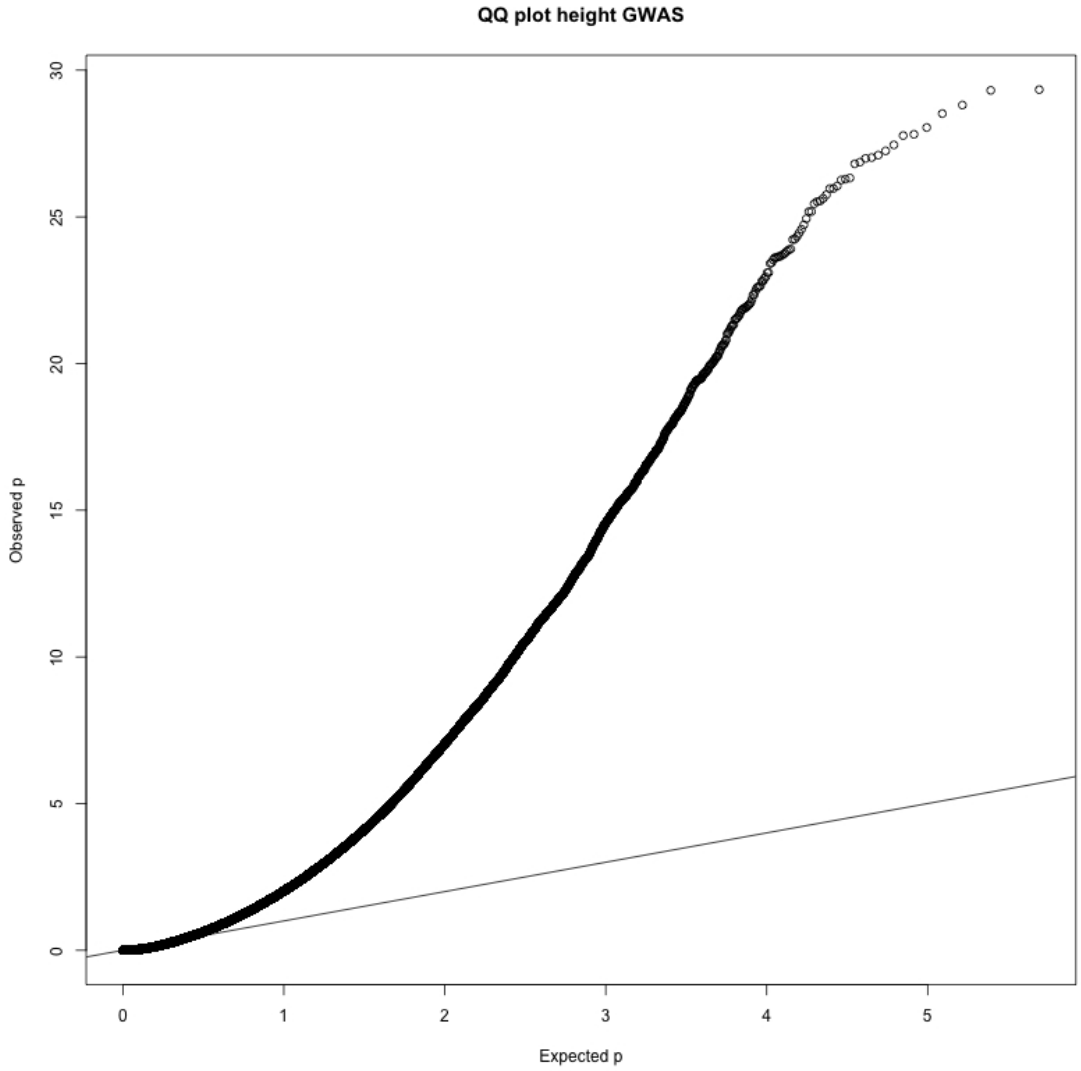


Figure 4.6: Quantile-quantile plot showing the distribution of -log10 p-values

GO term	Description	p-value	FDR
GO:0048856	Anatomical structure development	1.40E-06	0.003
GO:0009889	Regulation of biosynthetic process	7.10E-06	0.0034
GO:0009653	Anatomical structure morphogenesis	8.80E-06	0.0034
GO:0045449	Regulation of transcription	2.00E-05	0.0034
GO:0007275	Multicellular organismal development	1.10E-05	0.0034
GO:0019219	Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.60E-05	0.0034
GO:0019222	Regulation of metabolic process	1.90E-05	0.0034
GO:0006355	Regulation of transcription, DNA-dependent	1.90E-05	0.0034
GO:0032502	Developmental process	5.40E-06	0.0034
GO:0032501	Multicellular organismal process	1.00E-05	0.0034
GO:0051171	Regulation of nitrogen compound metabolic process	7.50E-06	0.0034
GO:0080090	Regulation of primary metabolic process	1.50E-05	0.0034
GO:0031326	Regulation of cellular biosynthetic process	1.90E-05	0.0034
GO:0051252	Regulation of RNA metabolic process	2.40E-05	0.0038
GO:0010556	Regulation of macromolecule biosynthetic process	2.70E-05	0.004
GO:0031323	Regulation of cellular metabolic process	3.70E-05	0.005
GO:0010468	Regulation of gene expression	4.00E-05	0.0052
GO:0060255	Regulation of macromolecule metabolic process	5.30E-05	0.0065
GO:0050789	Regulation of biological process	9.30E-05	0.011
GO:0048731	System development	0.00013	0.013
GO:0048869	Cellular developmental process	0.00013	0.013
GO:0048513	Organ development	0.00013	0.013
GO:0006350	Transcription	0.00016	0.014
GO:0006351	Transcription, DNA-dependent"	0.00016	0.014
GO:0065007	Biological regulation	0.00016	0.014
GO:0032774	RNA biosynthetic process	0.00018	0.015
GO:0050794	Regulation of cellular process	0.00022	0.018
GO:0032989	Cellular component morphogenesis	0.00039	0.03
GO:0022414	Reproductive process	0.00043	0.033
GO:0051239	Regulation of multicellular organismal process	0.00058	0.043
GO:0048646	Anatomical structure formation involved in morphogenesis	0.0006	0.043
GO:0000003	Reproduction	0.00069	0.048
GO:0003677	DNA binding	1.30E-05	0.0044
GO:0003682	Chromatin binding	8.50E-06	0.0044
GO:0030528	Transcription regulator activity	3.10E-05	0.0068
GO:0005488	Binding	9.40E-05	0.014
GO:0003700	Transcription factor activity	0.00011	0.014
GO:0004674	Protein serine/threonine kinase activity	0.0002	0.022
GO:0009508	Plastid chromosome	0.00015	0.03
GO:0000229	Cytoplasmic chromosome	0.00015	0.03

Table 4.4: Gene ontology enrichment results for the genes significantly associated with plant height in the maize FOAM landrace population

Chr	Position	Maize gene ID
1	172903811 -172907768	GRMZM5G852177
1	188114875 -188119970	GRMZM5G822593
1	202307514 -202313358	GRMZM2G491632
1	229897624 -229901197	GRMZM2G380650
1	84293878 -84299488	GRMZM2G040743
1	158473923 -158476588	GRMZM5G832908
1	170847950 -170852831	GRMZM2G052650
1	182100190 -182102156	GRMZM2G023591
1	206424076 -206426049	GRMZM2G454482
1	206794909 -206795232	$AC185611.3_FG001$
1	242140660 -242149635	GRMZM2G080462
1	244157841 -244159371	GRMZM2G320085
1	278605043 -278608272	GRMZM2G008072
2	12437551 -12451202	GRMZM2G313020
2	43339685 -43346040	GRMZM2G092107
2	87656247 -87657889	GRMZM2G154558
2	83267690 -83272352	GRMZM2G177575
2	118510311 -118512280	GRMZM2G034389
2	193501445 -193507266	GRMZM2G150503
2	196379519 -196383942	GRMZM2G151122
2	211280865 -211290331	GRMZM2G098859
2	26881250 -26885980	GRMZM2G062179
2	42525275 -42526930	GRMZM2G092621
2	144800491 -144803866	GRMZM2G092581
2	189936772 -189937946	GRMZM2G003947

2	205638604 -205642269	GRMZM2G073197
3	3994128 -3996072	GRMZM2G031432
3	4339914 -4345546	GRMZM2G045330
3	21563010 -21565627	GRMZM2G309897
3	159677432 -159679345	GRMZM2G030038
3	218893373 -218896627	GRMZM2G045318
3	4344526 -4345251	GRMZM2G346457
3	5786252 -5789252	GRMZM2G068352
3	22737686 -22738486	GRMZM2G360023
4	10900952 -10901590	$AC198403.3_FG001$
4	185866758 -185871668	GRMZM2G071288
4	230556470 -230569470	GRMZM2G038195
4	231896959 -231904271	GRMZM2G116079
4	179623979 -179624754	GRMZM2G374575
4	233800183 -233811237	GRMZM2G063909
5	187808501 -187809804	GRMZM2G159904
5	188652647 -188657031	GRMZM2G139837
5	198788740 -198790973	GRMZM2G004480
5	205783382 -205786772	GRMZM2G378665
5	75293187 -75312788	GRMZM2G009546
5	80715081 -80720509	GRMZM2G147800
5	193538315 -193539997	GRMZM2G165601
5	199721983 -199724840	GRMZM2G076313
6	98622638 -98628187	GRMZM2G074438
6	138182751 -138190937	GRMZM2G325804
6	141907720 -141909845	GRMZM2G088669

6	152177133 -152180071	GRMZM2G066448
6	163868874 -163870703	$AC195860.3_FG006$
6	91640314 -91641972	GRMZM2G128560
7	31086409 -31087079	GRMZM2G000221
7	133369333 -133372849	GRMZM2G103783
7	135573278 -135584080	GRMZM2G130959
7	136261560 -136272782	GRMZM2G097059
7	146491263 -146498155	GRMZM2G032003
7	161658285 -161664104	GRMZM5G813007
7	7414204 -7416229	GRMZM2G044498
7	72242504 -72287055	GRMZM2G013794
7	135508099 -135516934	GRMZM2G059225
8	21866130 -21869045	GRMZM2G333454
8	21877273 -21878956	GRMZM5G871018
8	111287695 -111290414	GRMZM2G333079
8	133178452 -133180059	GRMZM2G005552
8	164558359 -164563720	GRMZM2G117405
8	21874081 -21876830	GRMZM2G033283
8	22678378 -22682240	$AC187157.4_FG002$
8	124587539 -124589539	GRMZM2G071339
9	62247440 -62248096	GRMZM2G087350
9	115378515 -115380676	GRMZM2G069807
9	135853975 -135856576	GRMZM2G031637
9	139187765 -139188857	GRMZM2G305046
9	141434879 -141438278	GRMZM2G043854
9	151819897 -151823502	GRMZM2G010490

9	24485726 -24487699	GRMZM2G000932
9	62257531 -62258248	GRMZM2G401997
9	62257357 -62261260	GRMZM2G102322
9	99670405 -99672262	GRMZM2G468439
9	102912593 -102928246	GRMZM2G373578
9	139195465 -139196413	GRMZM5G825759
9	139191908 -139193008	GRMZM2G305027
9	147958762 -147965556	GRMZM5G895554
10	25556763 -25566637	GRMZM2G054393
10	87133277 -87135674	GRMZM2G109062
10	147134529 -147139119	GRMZM2G124502
10	147140446 -147143263	GRMZM2G124495
10	147147635 -147151283	GRMZM2G124476
10	148419923 -148423736	GRMZM2G142757
10	54465302 -54466708	GRMZM2G092363
10	66168601 -66179083	GRMZM2G070575
10	112937715 -112938912	GRMZM2G368047
10	141945134 -141951125	GRMZM2G031529

Table 4.5: Overlapping genes significantly associated with plant height in the maize FOAM and NAM panels

Chr	Position	Maize gene ID
1	189191520-189193031	GRMZM2G009117
1	241259218-241270285	GRMZM2G012611
1	243768555-243784387	GRMZM2G108463
2	197683480-197686929	$AC185415.3_FG005$
2	232964582-232968333	GRMZM2G082931
3	143622883-143624015	GRMZM2G010351
3	101146863-101147750	GRMZM2G013613
3	180041545-180043505	GRMZM2G059453
3	101735928-101737972	GRMZM2G062531
3	141261688-141325830	GRMZM2G073584
3	201299150-201303771	GRMZM2G079727
3	83042235-83069098	GRMZM2G081380
3	98442862-98443856	GRMZM2G094081
3	80514284-80517540	GRMZM2G100321
3	138772336-138774109	GRMZM2G101221
3	143695995-143699027	GRMZM2G102174
3	84903832-84905743	GRMZM2G112579
3	50093119-50099618	GRMZM2G119894
3	5580132-5586653	GRMZM2G143235
3	82052898-82106476	GRMZM2G154532
3	81036356-81074995	GRMZM2G156033
3	143800760-143803983	GRMZM2G164502
3	199250604-199254280	GRMZM2G177929
3	143605403-143606581	GRMZM2G335046
3	143487289-143488936	GRMZM2G369799
3	101037267-101041120	GRMZM2G401179
3	100248087-100285775	GRMZM2G409893
3	4861529-4865955	GRMZM2G471089
3	77230882-77231627	GRMZM2G701571
3	79945097-79945441	GRMZM2G701577
3	9303333-9314353	GRMZM5G844173
3	83782336-83782827	GRMZM5G865071
4	235606045-235606666	GRMZM5G821983
6	141275230-141278793	GRMZM2G084806
6	44583423-44592593	GRMZM2G156956
7	135508099-135516934	GRMZM2G059225
8	131967282-131968572	$AC219006.2_FG006$
8	130850435-130856002	GRMZM2G094241
8	132194624-132201905	GRMZM2G474726
8	123509007-123513267	GRMZM2G479987
8	132044001-132047428	GRMZM2G700665

Table 4.6: Overlapping genes significantly associated with plant height in the maize FOAM landrace population and the NCRPIS panel

CHAPTER 5

DISCUSSION

Crop landraces harbor a significant amount of useful alleles, yet they remain untapped because of the significant linkage drag associating each good allele to hundreds of undesirable alleles. In maize, efforts like the Latin American Maize project (LAMP)[1] highlighted the great diversity present in landraces. LAMP also showed that characterization of genetic resources present in germplasm banks is time and resource intensive, with adaptive patterns in maize landraces adding a layer of complexity, making accurate evaluation and transfer of alleles challenging. In addition, the lack of cost-effective marker technologies at the time limited the identification of genes underlying adaptive versus target traits, limiting the subsequent efficient use of beneficial alleles. Recently, the availability of genotyping technologies has enabled the use of association mapping for rapid identification and selection for increased carotenoid content in maize[2]. Although a significant milestone, the diversity surveyed in this and similar projects has been limited to improved inbred lines, and deployment has been generally achieved for traits with low complexity. Here we suggest new approaches for identifying useful alleles in landraces with potential application to several traits in maize and other species.

We characterized a panel of 4,500 individuals, each representative of one accession (population) from CIMMYT's maize gene bank. Together, these individuals represent a comprehensive panel of landraces encompassing the diversity from 35 countries in Latin America, which are grouped into 3 adaptation classes according to altitude. We genotyped those individuals for close to one million markers, representing the largest characterization effort in maize landraces to

date. For this panel, we explored linkage disequilibrium, population differentiation, and significance of known introgressed regions in maize.

Over the past several decades, cytological work has shown the presence of large structural variation in the form of chromosomal inversions segregating across maize populations[3]. Furthermore, recent work in inbred lines[4] shows that the recombination landscape of the maize genome is very complex, with higher recombination frequency in genic regions, and lower near centromeres. In addition, recent analyses show the important role of introgressions from teosinte into maize, in particular the important genetic contribution of *Zea mays* ssp *mexicana* and its effect on maize highland adaptation[5]. In contrast to such inter-species gene flow, important adaptive barriers limit gene flow within maize populations, as evidenced by the altitude-driven adaptation patterns observed in landraces[1]. We estimated an LD-like statistic using the non-phased SNP markers and we observed that 128 Mb of the maize genome display high LD. We observed a pattern consistent with lower recombination in the pericentromeric regions, with half a dozen high-LD regions in the chromosome arms. We also estimated the population differentiation index F_{ST} among the three adaptation classes, and observed that 90% of the genome remains with very low differentiation (F_{ST}=0.058). By comparing the regions with high-LD, high-F_{ST} and the 118 Mb reported as introgressions[5] we observed a significant overlap: 26.9% of the high F_{ST} regions are contained within a region reported as introgression from *Zea mays* ssp. *mexicana*; 99 Mb being introgressions with significantly increased in LD. The overlap between the *Zea mays* ssp. *mexicana* introgressions corresponds to the centromeres 1,5,6, and 10, as well as the chromosomal inversion INV4, 2 regions on chromosome 9, and 1 on chromosome 2. Together, these results show that the segregation of alleles in landraces is

tied to genome structure. Adaptive structural introgressions increase LD in regions that would otherwise experience high recombination. Furthermore centromeres, which generally show low recombination, also show evidence of introgressions in 4 out of the 10 maize chromosomes, displaying also significant differentiation across adaptation classes. This genomewide increased LD in landraces suggests that even with a very large effective population size, several regions of the maize genome experience inefficient selection against deleterious mutations, and at least one introgressed high-LD centromere has been reported previously with effect on heterosis[6]. The mosaicism of the maize genome, combining the remnants of the ancient polyploidy event, the modern segregation of adaptive structural variants, and the inefficient recombination around all the centromeres, shows the potential for technologies relying on faster and more precise transfer of useful alleles across populations of maize, and possibly teosintes, to significantly speed improvement efforts, as well as purging genetic load. This will probably be achieved as a combination of genomic selection aided by genome editing, with association results providing the allelic hypotheses.

In addition to genome-level analyses, we implemented a novel population design using the landrace individuals to directly dissect the genetic basis of complex traits, called F-One Association Mapping population. This design consists of crossing each individual with one hybrid of matching adaptation, therefore capturing 2 gametes per population. Evaluation is then performed on F1 progeny for specific traits, and phenotypic values are used for genome wide association. Using this design, we explored the genetic architecture of flowering time, plant height, and ear rot infection. For flowering time, we showed significant overlap at candidate genes, as well as a significant contribution of

large structural variation affecting flowering time in maize, including the adaptive inversion introgression INV4. Around 880 genes are involved in male or female flowering time, with an overlap of 72% across both traits. We showed that genomic prediction accuracy using the most significant markers 880 SNPs is equivalent to that of 30,000 markers, showing the potential for using the association results directly in breeding efforts. For plant height, we showed significance at one key domestication gene, *grassy tillers1*, which modifies plant architecture and apical dominance. In addition, we showed a significant enrichment at genes involved in regulation of biological processes, as well as several regions containing homologs known to be involved in the regulation of hormonal pathways in *Arabidopsis*. In contrast to flowering time, prediction accuracies for plant height remained low even when using all segregating markers. When comparing the minor allele frequencies of both traits, we observed a significant enrichment at low frequency polymorphisms, with height having the most pronounced effect.

Our FOAM approach has several advantages over other population designs. In contrast to mapping populations including biparental and Nested Association Mapping, it allows the sampling of diversity from thousands of populations, and it does not require the generation of higher filial generations after the F1, therefore speeding the process from sampling to association to deployment. Furthermore, unlike per se association mapping on landrace individuals, the replication of alleles across progeny and trials increases our ability to estimate the genetic effect of the parental landrace gametes. However, the nesting within adaptation necessary for accurate evaluation also limits evaluation of target traits outside the adaptive class of the landrace, and the lack of balanced replication across trials does not allow for accurate estimation of geno-

type by environment interaction effects. Furthermore, important trade-offs between FOAM and structured populations include decreased statistical power at minor allele frequencies less that 1%, with alleles probably affecting height on that fraction of the frequency spectrum. In addition, our current design does not allow the estimation dominance effects. Using inbred lines instead of hybrids as parental lines for the F1 progeny could help tackle this in the future, as long as sufficient progeny can be derived from those types of crosses, and balanced replication could help estimate genotype by environment interaction effects.

We also implemented a complementary approach to map genes likely involved in adaptation to abiotic stress. By utilizing the passport information on the landrace accessions, we are able to use the sampling location to infer climate and soil characteristics for the original populations. Those populations, having evolved in those environmental conditions through several generations, are enriched in alleles beneficial to growing under a variety of stresses. By performing association between genetic variation and environmental covariates, we are able to identify several regions of the genome associated to adaptation to differences in 10 climatic and 2 soil related traits. We showed that some traits have a significant contribution from introgression. In addition, we observed that the minor allele frequency of the alleles associated with environmental adaptation is high. Although high resolution was achieved, higher marker density will be necessary to identify the causal polymorphisms. Future work could include the cataloging of the alleles already present in various improved lines, as well as the movement of the untapped alleles from landraces into breeder's germplasm. This could help both validate their effect, study the physiological basis for the various adaptations and also could lead to programs matching adaptive alleles with changes in climatic conditions.

The major goals of breeding efforts are to guarantee sufficient and nutritious food supply. In face of climate change, the development of new varieties must also incorporate the ability to withstand the effects of extreme weather events, which are likely to increase in the next century. Characterizing and deploying the useful alleles from traditional varieties and wild species of domesticated animal and plant species will become a key tool for developing the varieties of the future. This can be achieved through a combination of new experimental designs, high density genotyping, high throughput phenotyping, and genome editing technologies.

5.1 References

1. Salhuana, W., Jones, Q. & Sevilla, R. The Latin American Maize Project: Model for rescue and use of irreplaceable germplasm. Diversity (1991).

2. Harjes, C. E. et al. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319, 330333 (2008).

3. McClintock, B., Yamakake, T. A. K. & Blumenschein, A. Chromosome constitution of races of maize: its significance in the interpretation of relationships between races and varieties in the Americas. (Colegio de Postgraduados Mexico, 1981).

4. Rodgers-Melnick, E. et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. U. S. A. 112, 38233828 (2015).

5. Hufford, M. B. et al. The genomic signature of crop-wild introgression in maize. PLoS Genet. 9, e1003477 (2013).

6. Stuber, C. W., Lincoln, S. E., Wolff, D. W., Helentjaris, T. & Lander, E. S. Iden-

tification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132, 823839 (1992).

APPENDIX A

EXPLORING THE POTENTIAL FOR FINDING SOURCES OF RESISTANCE TO FUSARIUM EAR ROT AMONG MAIZE LANDRACES

Ear rot is a group of severe and common diseases affecting maize, produced by several fungi species associated, which in turn produce mycotoxins that are toxic upon animal or human consumption[1]. From a breeding perspective, disease resistance has yet to be achieved through improvement. An important source for disease resistance in general are landraces, however strategies for evaluation and identification of markers associated with resistance on large numbers of accessions are necessary. Here we describe an effort to identify resistance alleles through evaluation of the maize landrace FOAM population.

In total 2 trials were conducted in 2012, and inoculation was performed with 2 different Fusarium species, Fusarium monilioiswa and Fusarium verticillioides. Inoculation is conducted 5-7 days after emergence of female flowers (silking) on the primary ear. A spore concentration of 5 x 10^6 spores/ml was prepared just before inoculation by diluting a concentrated fungal spore stock solution in distilled water. Inoculum was delivered using a plank with a nail inserted into it, and a swab. Inoculation is done by dipping the inoculating apparatus into the spore suspension to soak the foam rubber with spore solution, and then punching a hole into the middle of the ear, thereby depositing inoculum into the wounds. After inoculation, treated ears are marked with permanent marker pen or color in the outer surface of ears to identify inoculated ears during harvest. Harvest of inoculated ears per plot is done manually at physiological maturity. A visual score from levels 1-7 was used and

[0]For this appendix, my contribution encompassed the data analyses using the genotypic data, the genome wide association model and subsequent gene level analyses

percentage of infection was estimated using following formula: $\%Infection = [(1nx0) + (nx10) + (nx25) + (nx50) = (nx75) + (nx100)]/totalnumberofears)]$, where n is the total number of ears in each grade. Breeding values were estimated using the percentage Infection (Figure A.1). For Fusarium monilioiswa, there was no phenotypic variance, therefore only breeding values for Fusarium verticillioides were used as response variables for a Genome Wide Association. GWA was performed using a generalized linear model with a main effect for flowering date and 10 multidimensional scaling weights to account for population structure. There was one SNP with significant association after accounting for multiple testing (Figure A.2), and corresponded to the gene GRMZM2G086088, whose *Arabidopsis* homolog is annotated as a (UBC9) ubiquitin conjugating enzyme 9, which participates in protein sumoylation, a type of protein regulatory modification that has been implicated in several processes including defense response[2].

Landraces are an important source of beneficial alleles. Genome wide association has been used to identify alleles relevant for breeding objectives, and guide their use through marker assisted selection for example for biofortification of maize[3]. We performed genome wide association for percentage infection of Fusarium verticillioides on inoculated maize landrace trials. We detected significant association with one gene on chromosome 8, GRMZM2G086088. Further work will help establish the nature of the association, and if proteins related to the sumoylation pathway can help increase resistance to pathogens.

A.1 References

1. Munkvold, G. P. in Epidemiology of Mycotoxin Producing Fungi 705713 (Springer Netherlands, 2003).

2. Miura, K., Jin, J. B. & Hasegawa, P. M. Sumoylation, a post-translational regulatory process in plants. Curr. Opin. Plant Biol. 10, 495502 (2007).

3. Harjes, C. E. et al. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319, 330333 (2008).

A.2 Figures

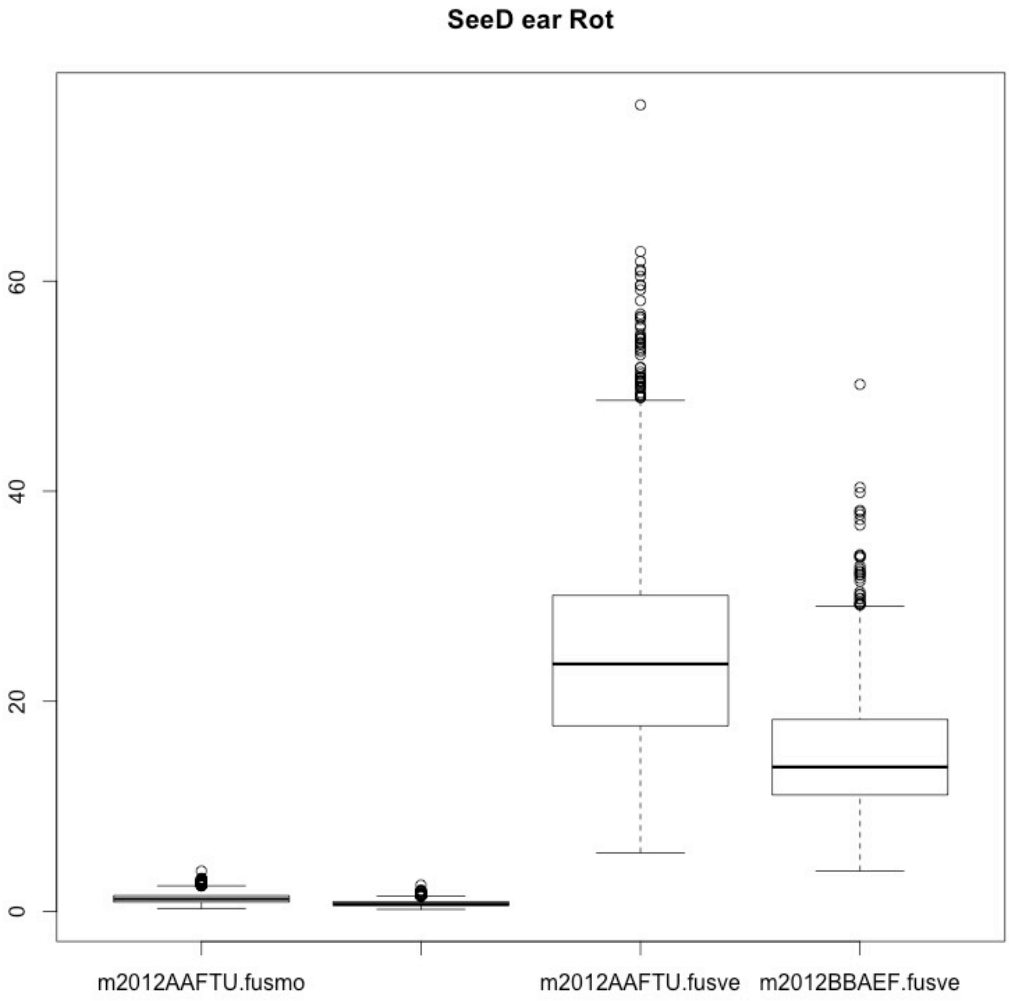


Figure A.1: Distribution of percentage infection for Fusarium inoculated trials

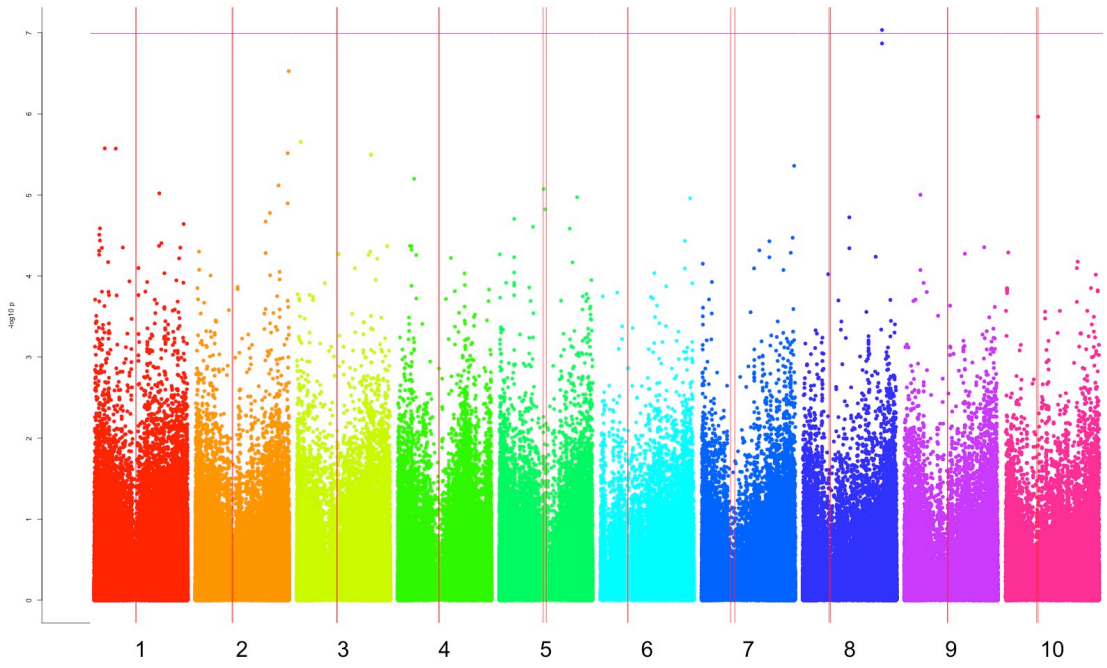


Figure A.2: Manhattan plot for Fusarium verticillioides percentage infection

APPENDIX B

INTEGRATION OF CONTROLLED POPULATIONS AND ASSOCIATION MAPPING TO SCORE CYTOLOGICAL FEATURES IN THE MAIZE GENOME

Understanding the mechanisms of normal and atypical inheritance is fundamental for the study of genetics. Meiotic drive is an irregular type self-promoted segregation of specific alleles during meiosis, and has emerged on several species independently, including Drosophila species[1], mouse[2], and maize[3], with each case using a different molecular mechanism. In maize, drive acts during female meiosis and is relevant for all chromosomes at regions of heterochromatic repeats called chromosomal knobs. Knobs are large heterochromatic segments composed and defined by their constituting repeats: 180-bp, TR-1 or composite. Knobs display presence/absence variation, as well as size variation, and are found in all maize chromosomes. The relevant feature for the drive system, named Abnormal 10 (Ab10), is a variant of chromosome 10 characterized by having the largest known chromosomal knob, as well as the genes responsible for the drive effect[4]. This system has had a significant effect on the current knob composition in the maize genome[5]. Currently, the only experimental strategy to determine the allele for Ab10 carried by an individual is to perform cytological identification. Here, we describe a potential method to score the Ab10 allele from individuals using genotyping by sequencing markers. By looking at the segregation in mapping populations, relevant markers were identified and used to infer the corresponding haplotype on unrelated individuals. By using the Ab10 haplotype, we identify a putative region that displays association consistent with current meiotic drive coupled to Ab10 segregation in maize

[0]For this appendix, my contribution encompassed the association studies

landraces.

In order to find the relevant SNPs to score Ab10, we used a panel of 91 self crossed individuals segregating for both Ab10-I and Ab10-II in a W23 background which was scored for R, a gene 1 cM away from Ab10 with r being linked with N10. We performed a genome wide association for the R state using a Generalized Linear Model with 1 Principal Components with the software TASSEL and found a few SNPs associated with R on chromosome 10 (Figure B.1). We looked at one megabase around the most significant SNP from the association to infer Ab10 clusters based on r status. Because of its effect increasing linkage disequilibrium (LD), we are able to infer haplotypes using multidimensional scaling. There were 4 samples that did not cluster according to the R status using this method: 5AB10IIr, 16AB10IR, 9*AB10IIR, 17*AB10IR. (Figure B.2) Because R is 1 cM from the beginning of the non-recombining haplotype region, and given these were self pollinated individuals and recombination is happening in both male and female, those samples are consistent with the 4 expected recombinants. We used the haplotype clustering to infer in the landraces the window of the genome with clustering consistent with same region (Figure B.3). We then used the corresponding Ab10 clusters as response variables in a genome wide scan accounting for population structure, and identified several regions in the genome with significant association with Ab10. In particular, a region on chromosome 4 was highly significant across a large region, consistent with meiotic drive of markers linked with a chromosomal knob (Figure B.5).

Recent studies point at the presence of Ab10 across Zea species, with a unique allele in low frequency in maize, compared to at least 2 more alleles segregating at higher frequency in teosintes[6]. Identifying the segregation of

Ab10 however remains limited by the time and resource intensive process of karyotypic characterization. We showed that by using publicly available high density genotypic data from maize landraces it is possible to predict the segregation of the Ab10 allele in those samples. We observed a segregation in low frequency across the Americas, consistent with previous observations (Figure B.4). In particular, by using the scored segregation of the Ab10 allele, we find a region on chromosome 4 with significant association with chromosome 10 allele, consistent with the region on chromosome 4 showing the signature drive around a chromosomal knob in the presence of Ab10. A few regions of the genome show a similar pattern, therefore it would be necessary to validate the observation with additional experiments. This work however shows that incorporating information from unrelated mapping populations can help unveil the presence of otherwise unscored genomic features in high-density genotyped individuals. Incorporating newly available genomic resources, for example long read sequencing data, will allow to better score Ab10 and other cytological features, which in turn will allow deeper understanding of the evolutionary and phenotypic consequences of such variants in maize and other species.

B.1 References

1. Gershenson, S. A New Sex-Ratio Abnormality in DROSOPHILA OBSCURA. Genetics 13, 488507 (1928).

2. Silver, L. M. The peculiar journey of a selfish chromosome: mouse t haplotypes and meiotic drive. Trends Genet. 9, 250254 (1993).

3. Rhoades, M. M. Preferential Segregation in Maize. Genetics 27, 395407 (1942).

4. Rhoades, M. M. & Dempsey, E. The Effect of Abnormal Chromosome 10

on Preferential Segregation and Crossing over in Maize. Genetics 53, 9891020 (1966).

5. Buckler, E. S., 4th et al. Meiotic drive of chromosomal knobs reshaped the maize genome. Genetics 153, 415426 (1999).

6. Kanizay, L. B. et al. Diversity and abundance of the abnormal chromosome 10 meiotic drive complex in Zea mays. Heredity 110, 570577 (2013).

B.2 Figures

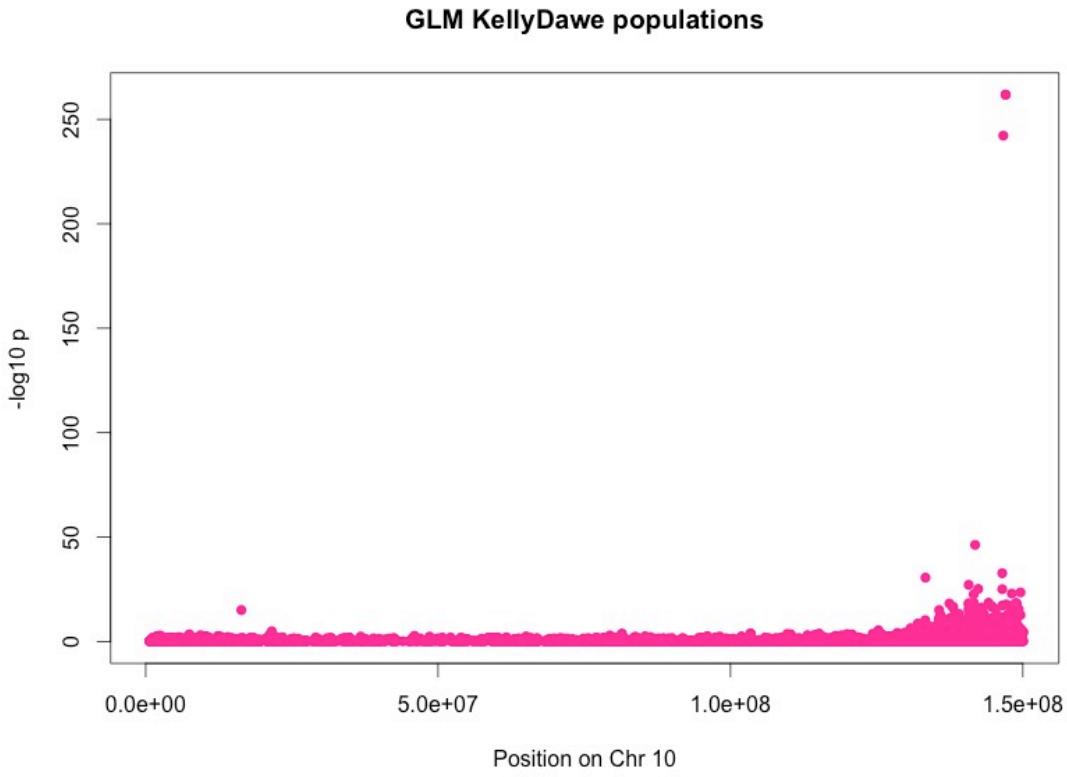


Figure B.1: Local Manhattan plot for chromosome 10 using r scored value

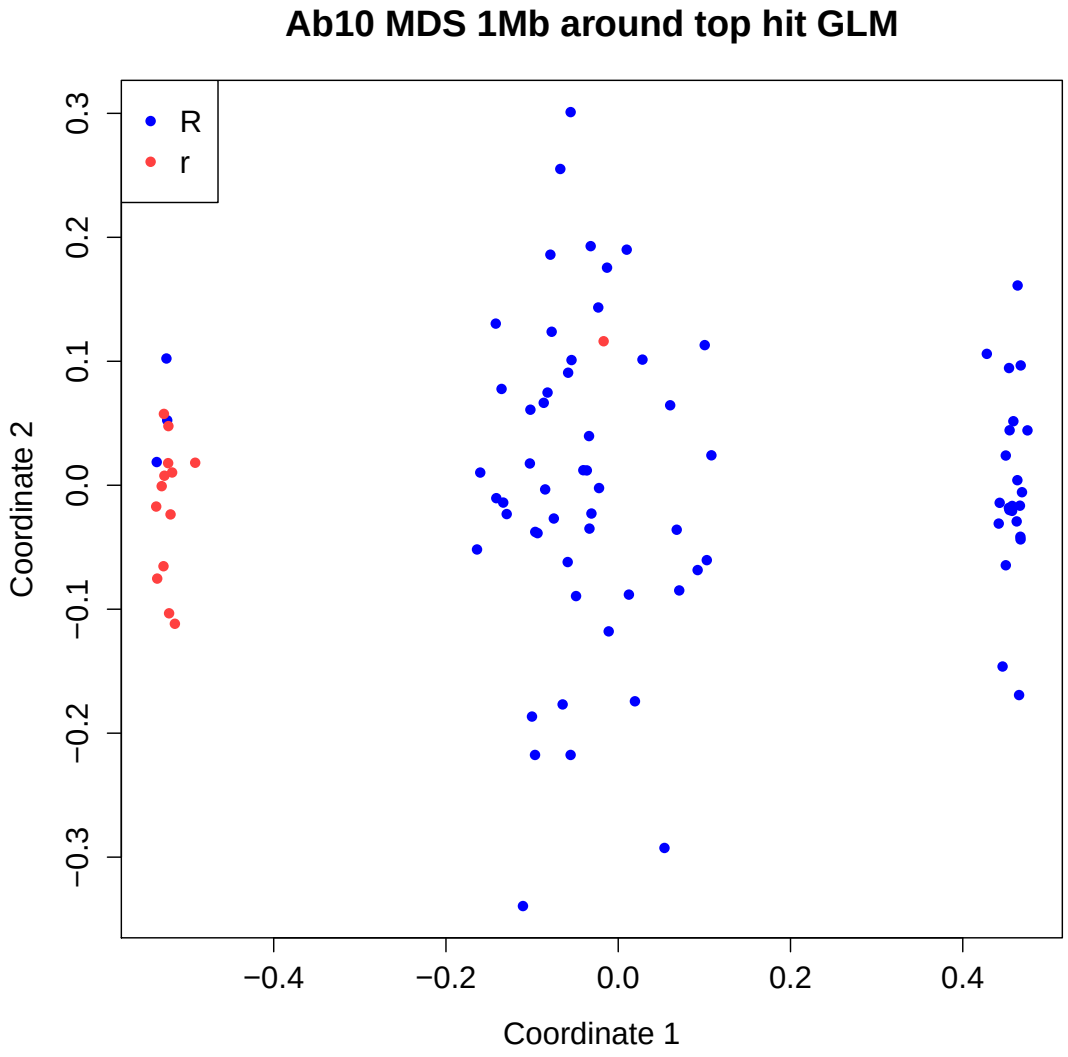


Figure B.2: Multidimensional Scaling around top hit for scored r value

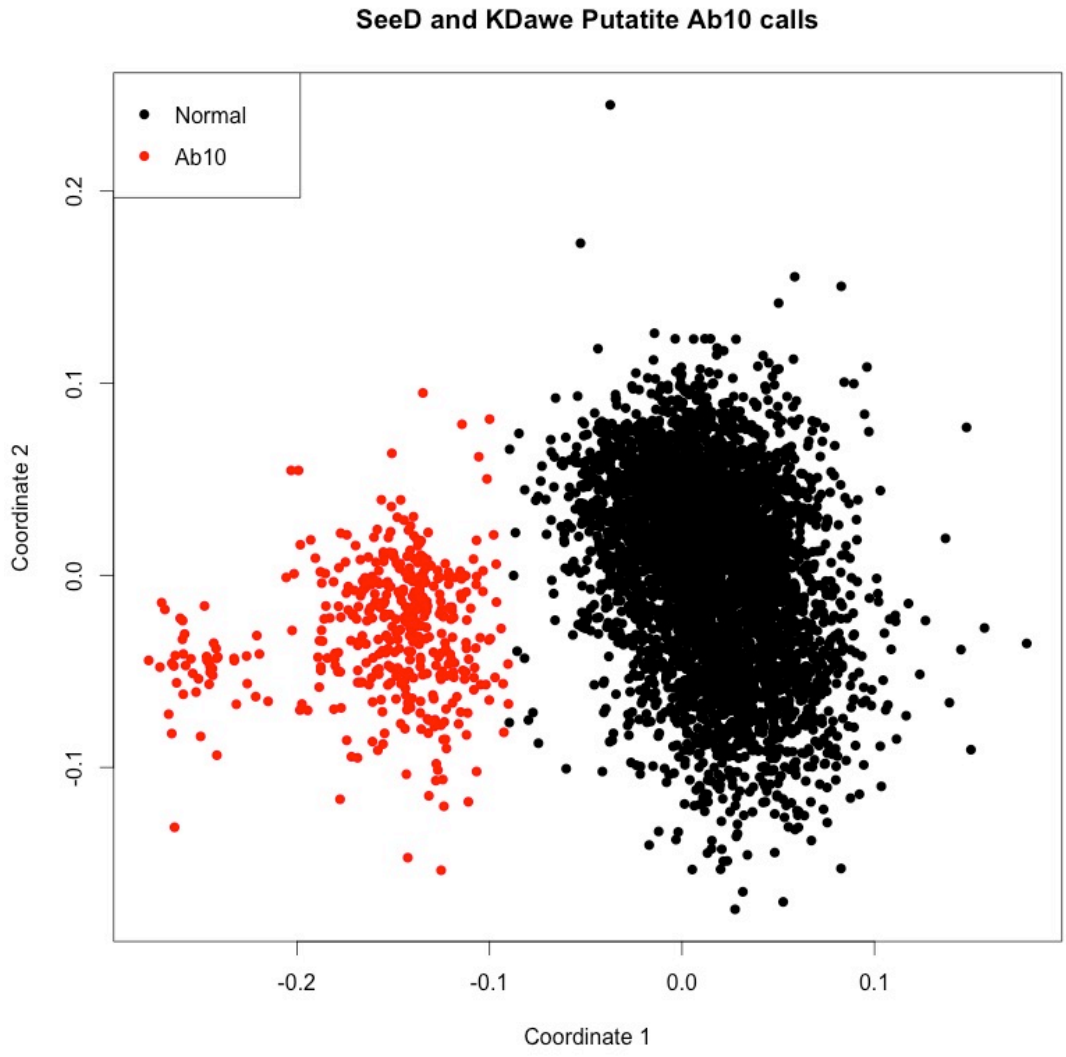


Figure B.3: Multidimensional Scaling around top hit for r in the FOAM landrace parents

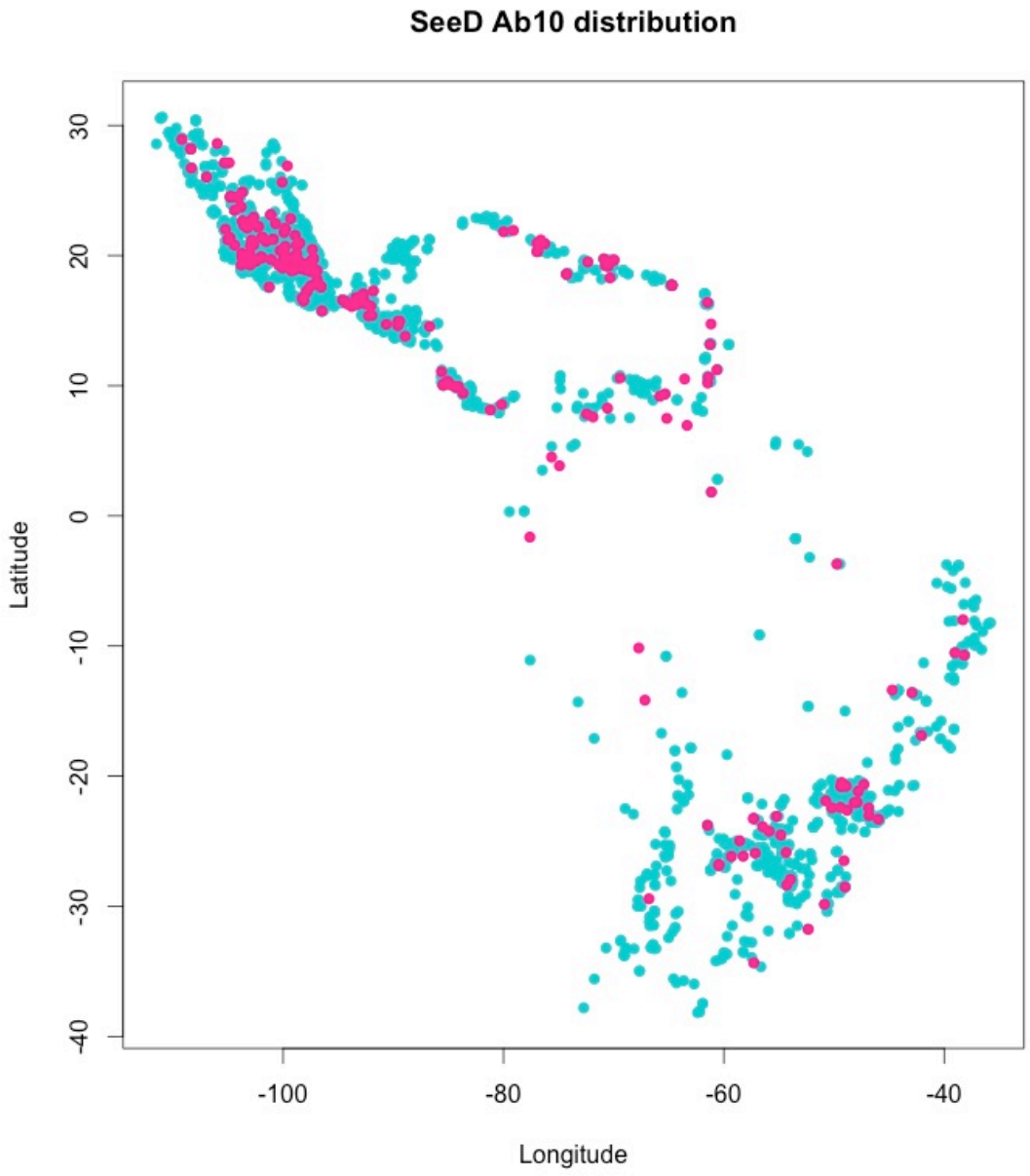


Figure B.4: Sampling location from Ab10 carrying landrace accessions

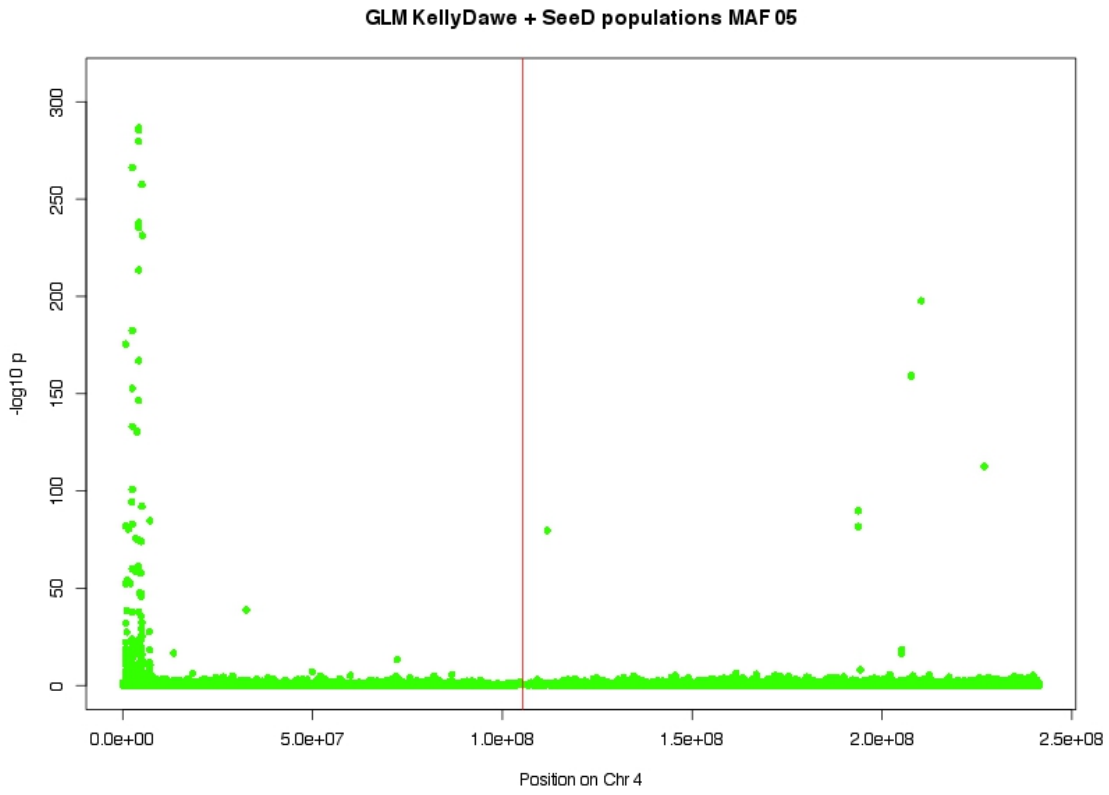


Figure B.5: Local Manhattan plot on chromosome 4 displaying putative knob positions